A large-scale service system with packing constraints: Minimizing the number of occupied servers

Alexander L. Stolyar
Bell Labs, Alcatel-Lucent
600 Mountain Ave., 2C-322
Murray Hill, NJ 07974
stolyar@research.bell-labs.com

Yuan Zhong
University of California
465 Soda Hall, MC-1776
Berkeley, CA 94720
zhyu4118@berkeley.edu

ABSTRACT

We consider a large-scale service system model proposed in [14], which is motivated by the problem of efficient placement of virtual machines to physical host machines in a network cloud, so that the total number of occupied hosts is minimized. Customers of different types arrive to a system with an infinite number of servers. A server packing *configuration* is the vector $\boldsymbol{k} = \{k_i\}$, where k_i is the number of type-i customers that the server "contains". Packing constraints are described by a fixed finite set of allowed configurations. Upon arrival, each customer is placed into a server immediately, subject to the packing constraints; the server can be idle or already serving other customers. After service completion, each customer leaves its server and the system.

It was shown in [14] that a simple real-time algorithm, called *Greedy*, is asymptotically optimal in the sense of minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}^{1+\alpha}$ in the stationary regime, as the customer arrival rates grow to infinity. (Here $\alpha > 0$, and $X_{\boldsymbol{k}}$ denotes the number of servers with configuration $\boldsymbol{k}$.) In particular, when parameter α is small, *Greedy* approximately solves the problem of minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$, the number of occupied hosts. In this paper we introduce the algorithm called *Greedy with sublinear Safety Stocks (GSS)*, and show that it asymptotically solves the exact problem of minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$. An important feature of the algorithm is that sublinear safety stocks of $X_{\boldsymbol{k}}$ are created automatically – when and where necessary – without having to determine *a priori* where they are required. Moreover, we also provide a tight characterization of the rate of convergence to optimality under *GSS*. The *GSS* algorithm is as simple as *Greedy*, and uses no more system state information than *Greedy* does.

Categories and Subject Descriptors

[**Network Services**]: Cloud Computing; [**Probability and Statistics**]: Markov Processes, Queueing Theory, Stochastic Processes; [**Design and Analysis of Algorithms**]: Approximation Algorithms Analysis, Packing and Covering Problems

General Terms

Algorithms, Performance, Theory

Keywords

Multi-dimensional Bin Packing, Infinite-Server System, Markov Chain, Safety Stocks, Fluid Scale Optimality, Local Fluid Scaling

1. INTRODUCTION

We consider a service system model [14] motivated by the problem of efficient placement of virtual machines (VMs) to physical host machines (servers) in a data center (DC) [6]. A *service policy* decides to which server each incoming VM will be placed. We are interested in service policies that minimize the total number of occupied servers in the system. It is further desirable that the policy be simple, so that placement decisions are made in real time, and depend only on the current system state, but not on system parameters.

Consider the following description of a DC. It consists of a number of servers. While servers may potentially have different characteristics, in this paper we assume that they are all the same. More specifically, let there be N different types of resources (for example, type-1 resource can be CPU, type-2 resource can be memory, etc). For each $n \in \{1, 2, \ldots, N\}$, a server possesses amount $B_n > 0$ of type-n resource. I types of VMs arrive in a probabilistic fashion, and request services at the DC. Arriving VMs will be placed into the servers, occupying certain resources. More specifically, for $i \in \{1, 2, \ldots, I\}$, a type-$i$ VM requires amount $b_{i,n} > 0$ of type-n resource during service, where $n \in \{1, 2, \ldots, N\}$. Once a VM completes its service, it departs the system, freeing up corresponding resources. We assume that service times of different VMs are independent.

For each $i \in \{1, 2, \ldots, I\}$, let k_i be the number of type-i VMs that a server contains. Then the following *vector packing constraints* must be observed at all times. Namely, a server can contain k_i type-i VMs ($i \in \{1, 2, \ldots, I\}$) simultaneously if and only if

$$\sum_i k_i b_{i,n} \leq B_n, \tag{1}$$

for each $n \in \{1, 2, \ldots, N\}$. In this case, the vector $\boldsymbol{k} = (k_1, \ldots, k_I)$ is called a *server configuration*.

The model considered in this paper is similar to the DC described above, but different in the following two aspects.

1. While vector packing constraints (cf. Eq. (1)) arise naturally in the context of VM placement, we make the more general assumption of so-called *monotone* packing constraints (cf. Section 2.1) in our model.

2. We consider a system with an infinite number of servers, where incoming VMs will be immediately placed into a server. For large-scale DCs, the number of servers is not a bottleneck, hence an infinite-server system reasonably approximates such DCs.

We would also like to remark that an important assumption of our model is that the service requirement of a VM is not affected by potentially other VMs occupying the same server. This is a reasonable modeling assumption for multi-core servers, for example.

There can be different performance objectives of interest. For example, we may be interested in minimizing the total energy consumption [6], or maximizing system throughput [9]. In this paper, we are interested in minimizing the total number of occupied servers. These objectives are different but related. For example, by switching off idle servers, or keeping them in stand-by mode, we can reduce energy consumption by minimizing the number of occupied servers.

In the main results of the paper, we introduce the policy called *Greedy with sublinear Safety Stocks (*GSS*)*, and show that it asymptotically minimizes the total number of occupied servers in steady state, as the input flow rates of VMs grow to infinity. *GSS* is a simple policy that makes placement decisions in real time, and based only on the current system state. Informally speaking, *GSS* places incoming VMs in a way that greedily minimizes a Lyapunov function, which asymptotically coincides with the total number of occupied servers. *GSS* maintains non-empty safety stocks at every server configuration $\boldsymbol{k}$ whenever $X_{\boldsymbol{k}}$ becomes "too small", so as to allow flexibility on VM placement. In other words, under *GSS*, there is a non-zero number of servers of every configuration, so that an incoming VM can potentially be placed into a server with any configuration. These safety stocks correspond to the discrepancy between the Lyapunov function and the total number of occupied servers, and grow "sublinearly" with the input flow rates. We also provide a characterization of the rate of convergence to optimality under *GSS*, which is tighter than the conventional fluid-scale convergence rate.

1.1 Related Works

In this section, we discuss related works, and put our results in perspective.

The most closely related work is [14], where the model considered in this paper was proposed, and a related problem was studied. In both this paper and [14], the asymptotic regime of interest is when the input flow rates grow to infinity, and the system is considered under the *fluid scaling*, i.e., when the system states are scaled down by the input flow rates. In [14], the problem of interest is minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}^{1+\alpha}$, where $\alpha > 0$, and $X_{\boldsymbol{k}}$ is the number of occupied servers with configuration $\boldsymbol{k}$. A simple policy called *Greedy* was introduced, which asymptotically minimizes the sum $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}^{1+\alpha}$, for any $\alpha > 0$, in the stationary regime. Policies *Greedy* and *GSS* differ in two important aspects. First, they try to minimize different objectives – $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}^{1+\alpha}$ ($\alpha > 0$) and $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$, respectively. When $\alpha > 0$ is small, *Greedy* approximately solves the problem of minimizing the total number of occupied servers $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$, in the asymptotic regime where the input flow rates grow to infinity, and at the fluid scale. However, if minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$ is the "true" desired objective, $\alpha > 0$ need to be chosen carefully, depending on the system scale (input flow rates), which may be difficult to do. Therefore, we believe that asymptotically solving the exact problem of minimizing $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$ is of substantial interest. Moreover, the policy *GSS* proposed in this paper is as simple as *Greedy*, and uses no more system state information than *Greedy* does. Second, at a technical level, to prove the asymptotic optimality of *Greedy*, [14] considered only the fluid scaling and the corresponding fluid limits. In this paper, to prove the asymptotic optimality of *GSS*, it is no longer sufficient to consider the fluid-scale system behavior alone; a *local fluid scaling* is also considered, needed to study the dynamics of safety stocks. In addition, this allows us to derive a tighter characterization of the rate of convergence to optimality under *GSS*, as opposed to the fluid-scale convergence shown in [14] for *Greedy*.

On a broader level, the model considered in this paper is related to the vast literature on classical stochastic bin packing problems. In a bin packing system, random-sized items arrive, and need to be placed into finite-sized bins. The items do not leave or move between bins, and a typical objective is to minimize the number of occupied bins. A packing problem is *one-dimensional* if sizes of the items and bins are captured by scalars, and *multi-dimensional* if they are captured by vectors. Problems with the multi-dimensional packing constraints (1) are called *vector packing*. For a good review of one-dimensional bin packing, see for example [2], and see for example [1] for a recent review of multi-dimensional packing. In bin packing *service* systems, items (customers) arrive at random times to be placed into a bin (server), and leave after a random service time. The servers can process multiple customers as long as packing constraints are observed. Customers get queued, and a typical objective of a packing algorithm is to maximize system throughput. (See for example [4] for a review of this line of work.) Our model is similar to the latter systems, except there are multiple bins (servers) – in fact, an infinite number in our case. Models of this type are more recent (see for example, [8, 9]). [8] addresses a joint routing and VM placement problem, which in particular includes packing constraints. The approach of [8] resembles Markov Chain algorithms used in combinatorial optimization. [9] considers maximizing throughput of a queueing system with a finite number of bins (servers), where VMs can wait for service. Very recently, [7] has new results on the classical one-dimensional online bin packing; it also contains heuristics and simulations for the corresponding system with item departures, which is a special case of our model.

As mentioned earlier, we consider the asymptotic regime where the input flow rates scale up to infinity. In this respect, our work is related to the (also vast) literature on queueing systems in the *many servers* regime. (See e.g. [12] for an overview. The name "many servers" reflects the fact that the average number of occupied servers scales up to infinity as well, linearly with the input flow rates.) However, packing constraints are not present in earlier works (prior to [14]) on the many servers regime, to the best of

our knowledge.

The idea of maintaining sublinear safety stocks to increase system flexibility, and hence avoid "resource" starvation – the approach taken by *GSS*, the policy proposed in this paper – has also appeared in other works. For example, see [10] and the references therein for an overview. However, to the best of our knowledge, the following feature of *GSS* is novel, and has not appeared in algorithms proposed in earlier works. Namely, *GSS* creates safety stocks *automatically*, in the sense that it does not require *a priori* knowledge of the subset of configurations for which the sublinear safety stocks need to be maintained. As a result, *GSS* does not require any *a priori* knowledge of the system parameters, because the safety stocks automatically adapt to parameter changes. We remark that the policy *Greedy* proposed in [14] also creates safety stocks, but they scale linearly with the input flow rates, whereas *GSS* creates sublinear safety stocks.

Finally, an overview of some resource allocation issues that arise from VM placement in the context of cloud computing can be found in [6].

1.2 Organization

The rest of the paper is organized as follows. In Section 1.3, we introduce the notation and conventions adopted in this paper. The precise model and main results are described in Section 2. The model is introduced in Section 2.1. Here we describe two versions of the model, the closed and open system. In Section 2.2, we describe the asymptotic regime of interest. The *GSS* policy is described in Section 2.3, and the main results, Theorems 6 and 7, are stated in Section 2.4, for the closed and open system, respectively. Sections 3 and 4 are devoted to proving Theorems 6 and 7, respectively. A discussion of the results in this paper and some future directions is provided in Section 5.

1.3 Notation and Conventions

Let $\mathbb{R}$ be the set of real numbers, and let $\mathbb{R}_+$ be the set of nonnegative real numbers. Let $\mathbb{Z}$ be the set of integers, let $\mathbb{Z}_+$ be the set of nonnegative integers, and let $\mathbb{N}$ be the set of natural numbers. $\mathbb{R}^n$ denotes the real vector space of dimension n, and $\mathbb{R}^n_+$ denotes the nonnegative orthant of $\mathbb{R}^n$. $\mathbb{Z}^n$ and $\mathbb{Z}^n_+$ are similarly defined. We reserve bold letters for vectors, and plain letters for scalars and sets. For a scalar x, let $|x|$ denote its absolute value, and let $\lceil x \rceil$ denote the largest integer that does not exceed x. For two scalars x and y, let $x \wedge y = \min\{x, y\}$, and let $x \vee y = \max\{x, y\}$. For a vector $\boldsymbol{x} = (x_i)_{i=1}^n \in \mathbb{R}^n$, let $\|\boldsymbol{x}\|$ denote its 1-norm, i.e., $\|\boldsymbol{x}\| = \sum_{i=1}^n |x_i|$. The distance from vector $\boldsymbol{x} \in \mathbb{R}^n$ to a set $U \subset \mathbb{R}^n$ is denoted by $d(\boldsymbol{x}, U) = \inf_{\boldsymbol{u} \in U} \|\boldsymbol{x} - \boldsymbol{u}\|$. We use $\boldsymbol{e}_i$ to denote the i-th standard unit vector, with only the ith component being 1, and all other components being 0. For a set $\mathcal{N}$, let $\mathbf{1}_{\mathcal{N}}$ be the indicator function of $\mathcal{N}$. For a finite set $\mathcal{N}$, let $|\mathcal{N}|$ be its cardinality. For two sets $\mathcal{N}$ and $\mathcal{M}$, let $\mathcal{N} \backslash \mathcal{M}$ denote the set difference of $\mathcal{N}$ and $\mathcal{M}$, i.e., $\mathcal{N} \backslash \mathcal{M} = \{x \in \mathcal{N} : x \notin \mathcal{M}\}$. For a set $\mathcal{N} \subset \mathbb{R}^n$, let $\langle \mathcal{N} \rangle$ denote its convex hull, i.e., the set of all $\boldsymbol{x} \in \mathbb{R}^n$ such that there exist $\gamma_1, \ldots, \gamma_m \in \mathbb{R}_+$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathcal{N}$ with $\boldsymbol{x} = \sum_{j=1}^m \gamma_j \boldsymbol{v}_j$ and $\sum_{j=1}^m \gamma_j = 1$. Symbol $\to$ means ordinary convergence in $\mathbb{R}^n$, and $\Longrightarrow$ denotes convergence in distribution of random variables taking values in $\mathbb{R}^n$, equipped with the Borel σ-algebra. The abbreviation *w.p.1* means convergence *with probability 1*. We often write $x(\cdot)$ to mean the function (or random process) $\{x(t),\ t \geq 0\}$. We write iff as a shorthand for "if and only if", i.o for "infinitely often", LHS for "left-hand side" and RHS for "right-hand side". We also write WLOG for "without loss of generality", w.r.t for "with respect to", and u.o.c for "uniformly on compact sets".

Throughout this paper, if $x(\cdot)$ is a random process (which in most cases will be Markov), we will denote by $x(\infty)$ its random state when the process is in stationary regime; in other words, $x(\infty)$ is equal in distribution to $x(t)$ (for any t) when $x(\cdot)$ is stationary. We use the terms *steady state* and *stationary regime* interchangeably.

2. MODEL AND MAIN RESULTS

2.1 Infinite Server System with Packing Constraints

We consider the following infinite server system that evolves in continuous time. There are I types of customers, indexed by $i \in \{1, 2, \ldots, I\} \equiv \mathcal{I}$, and an infinite number of homogeneous servers. A server can potentially serve more than one customer simultaneously. We use $\boldsymbol{k} = (k_1, k_2, \ldots, k_I) \in \mathbb{Z}^I_+$, an I-dimensional vector with nonnegative integer components, to denote a *server configuration*. The general packing constraints are captured by the finite set $\bar{\mathcal{K}} \subset \mathbb{Z}^I_+$ of *feasible server configurations*. Thus, a server can simultaneously serve k_i customers of type i, $i \in \mathcal{I}$, iff $\boldsymbol{k} = (k_1, k_2, \ldots, k_I) \in \bar{\mathcal{K}}$. From now on, we drop the word "feasible", and simply call $\bar{\mathcal{K}}$ the set of server configurations.

In this paper, we assume that the set $\bar{\mathcal{K}}$ is *monotone*.

ASSUMPTION 1. *$\bar{\mathcal{K}}$ is* monotone *in the following sense. If $\boldsymbol{k} \in \bar{\mathcal{K}}$, and $\boldsymbol{k}' \in \mathbb{Z}^I_+$ has $\boldsymbol{k}' \leq \boldsymbol{k}$ component-wise, then $\boldsymbol{k}' \in \bar{\mathcal{K}}$ as well.*

A simple consequence of the monotonicity assumption is that $\mathbf{0} \in \bar{\mathcal{K}}$. We now let $\mathcal{K} = \bar{\mathcal{K}} \backslash \{\mathbf{0}\}$ denote the set of non-zero server configurations.

Vector Packing is Monotone. An important example of monotone packing is vector packing. Consider the vector packing constraints in (1). It is clear that if the server configuration $\boldsymbol{k} = \{k_1, \ldots, k_I\}$ satisfies (1), and if $\boldsymbol{k}' \leq \boldsymbol{k}$ component-wise, then $\boldsymbol{k}'$ also satisfies (1). On the other hand, not all monotone packing is vector packing. For example, when $I = 2$, $\bar{\mathcal{K}} = \{(0,0), (0,1), (0,2), (1,0), (2,0)\}$ is monotone, but is not described by vector packing constraints. In the sequel, we will only assume monotone packing in our model, and all our results hold under this general setting.

To exclude triviality, we also assume that for all $i \in \mathcal{I}$, $\boldsymbol{e}_i$ (the i-th standard unit vector) is an element of $\bar{\mathcal{K}}$.

As discussed in the introduction, we make the following important assumption in this paper. We assume that simultaneous services do *not* affect the service distributions of individual customers; in other words, the service time of a customer is unaffected by whether or not there are other customers served simultaneously by the same server. Let us also remark that ideally, we would like to consider an open system, where each arriving customer is immediately placed for service in one of the servers, and leaves the system after service completion. However, we will first consider a "closed" version of this open system. The reason is twofold. First, the analysis of the closed system is a stepping stone

to that of the open system, and illustrates the main ideas more clearly. Second, we will see shortly that the closed system can be used to model job migration in a cloud, and is therefore of independent interest.

Denote by $X_{\boldsymbol{k}}$ the number of servers with configuration $\boldsymbol{k} \in \mathcal{K}$. The system state is then the vector $\boldsymbol{X} = \{X_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K}\}$. By convention, $X_{\boldsymbol{0}} \equiv 0$ at all times.

Closed System. Here we describe the "closed" version of the model. Let $r \in \mathbb{N}$ be given. Suppose that there are in total r customers in the system, and no exogenous arrivals. For each $i \in \mathcal{I}$, we suppose that there are $\rho_i r$ customers of type i in the system at all times. This in particular implies that $\sum_{i\in\mathcal{I}} \rho_i = 1$. It is convenient to index the system by r its total number of customers, and we use $\boldsymbol{X}^r = (X^r_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K})$ to denote a system state. The system evolves as follows. Each customer is almost always in service, except at a discrete set of time instances, where it migrates from one server to another (possibly the same one), subject to the packing constraints imposed by $\bar{\mathcal{K}}$. For a customer, the time between consecutive migrations is called its *service requirement*. Thus, one can alternatively think of a customer as departing the system after its service requirement, and then immediately arriving to the system, to be placed into a server. For each i, we assume that the service requirements of type-i customers are i.i.d. exponential random variables with mean $1/\mu_i$, and that the service requirements are independent across different $i \in \mathcal{I}$. A (Markovian) *service policy* ("packing rule") decides to which server a customer will be placed after its service requirement, based only on the current system state $\boldsymbol{X}^r$. A service policy has to observe the packing constraints. Under any well-defined service policy, the system state at time t, $\boldsymbol{X}^r(t)$, is a continuous-time Markov chain on a finite state space. Hence, for each r, the process $\{\boldsymbol{X}^r(t),\ t \geq 0\}$ always has a stationary distribution.

Open System. In the open system, customers of type i arrive exogenously as an independent Poisson flow of rate $\lambda_i r$, where λ_i is fixed and r is a scaling parameter. Each arriving customer has to be placed for service immediately in one of the servers, subject to the packing constraints imposed by $\bar{\mathcal{K}}$. Service times of all customers are independent. Service time of a type-i customer is exponentially distributed with mean $1/\mu_i$. After a service completion, each customer leaves the system. If we denote $\rho_i = \lambda_i/\mu_i$, then in steady state, the average number of type i customers in the system is $\rho_i r$, and the average total number of customers is $\sum_i \rho_i r$. We assume, WLOG, that $\sum_i \rho_i = 1$ – this is equivalent to re-choosing the value of parameter r, if necessary. A (Markovian) *service policy* ("packing rule") in this case decides to which server an arriving customer will be placed, based only on the current system state. A service policy has to observe the packing constraints. Similar to the closed system, we let $X^r_{\boldsymbol{k}}(t)$ denote the number of servers with configuration $\boldsymbol{k}$ at time t in the rth system. However, for the policy that we will study, $\boldsymbol{X}^r(t) = (X^r_{\boldsymbol{k}}(t))_{\boldsymbol{k}\in\mathcal{K}}$ will not be a Markov process. We postpone the discussion of a complete Markovian description of the system and the existence of the associated stationary distribution to Section 2.3.2.

2.2 Asymptotic Regime

We are interested in finding a service policy that minimizes the total number of occupied servers in the stationary regime. The exact problem is intractable, so instead we consider asymptotically optimal service policies. For both the closed and open systems, the asymptotic regime of interest is when $r \to \infty$. Informally speaking, in this limit, the *fluid-scaled* system state satisfies a conservation law (cf. Eq. (4)), and the best that a policy can do is solving a linear program, subject to this conservation law. We now describe the asymptotic regime in more detail.

First, we defined the so-called *fluid scaling*. Recall that both the closed and open systems are indexed by r, and $\boldsymbol{X}^r(t)$ is the vector that denotes the numbers of servers at time t, in the rth system. The *fluid scaled* process is $\boldsymbol{x}^r(t) = \boldsymbol{X}^r(t)/r$. For each r, in the closed system, $\boldsymbol{X}^r(\cdot)$ has a (not necessarily unique) stationary distribution, so $\boldsymbol{x}^r(\cdot)$ also has a stationary distribution. We will see shortly that in an open system, $\boldsymbol{X}^r(\cdot)$ also has a stationary distribution (see Lemma 5). Denote by $\boldsymbol{X}^r(\infty)$ and $\boldsymbol{x}^r(\infty)$ the random states of the corresponding processes in a stationary regime. (Recall the convention in Section 1.3.)

We now argue that as $r \to \infty$,

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x^r_{\boldsymbol{k}}(\infty) \implies \rho_i, \text{ for all } i. \tag{2}$$

In a closed system, for each $i \in \mathcal{I}$, there are $\rho_i r$ customers of type i in the system at all times, so on all sample paths,

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x^r_{\boldsymbol{k}}(t) = \rho_i, \text{ for all } r, t \text{ and } i.$$

This implies that the same holds for $\boldsymbol{x}^r(\infty)$. In an open system, the total number of type-i customers is $\sum_{\boldsymbol{k}\in\mathcal{K}} k_i X^r_{\boldsymbol{k}}(\infty)$, in steady state. It is easy to see that, independent from the service policy, this quantity is a Poisson random variable with mean $\rho_i r$. Thus, as $r \to \infty$, $\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x^r_{\boldsymbol{k}}(\infty) \implies \rho_i$.

Now consider the following linear program (LP).

$$\text{Minimize} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} \tag{3}$$

$$\text{subject to} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, \quad \text{for all } i \in \mathcal{I}, \tag{4}$$

$$x_{\boldsymbol{k}} \geq 0, \quad \text{for all } \boldsymbol{k} \in \mathcal{K}. \tag{5}$$

Denote by $\mathcal{X}$ the set of feasible solutions to LP:

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|} : \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, i \in \mathcal{I}\}.$$

Then $\mathcal{X}$ is a compact subset of $\mathbb{R}_+^{|\mathcal{K}|}$. Let $\mathcal{X}^*$ denote the set of optimal solutions of LP, and let u^* denote its optimal value. In light of Eqs. (2) and (4), a service policy is asymptotically optimal if, roughly speaking, under this policy and for large r, $\sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}}(\infty) \approx u^*$ with high probability (cf. Theorems 6 and 7).

The following characterization of the set $\mathcal{X}^*$ by dual variables will be useful. The proof is elementary and omitted.

LEMMA 2. *$\boldsymbol{x} = (x_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{K}} \in \mathcal{X}^*$ iff $\boldsymbol{x}$ is a feasible solution of LP, and there exist $\eta_i \in \mathbb{R}$, $i \in \mathcal{I}$, such that*

(i) $\sum_{i\in\mathcal{I}} k_i\eta_i \leq 1$ for all $\boldsymbol{k} \in \mathcal{K}$, and

(ii) if $\sum_{i\in\mathcal{I}} k_i\eta_i < 1$, then $x_{\boldsymbol{k}} = 0$.

The following lemma relates the distance between a point $\boldsymbol{x} \in \mathcal{X}$ and the optimal set $\mathcal{X}^*$ to the objective value of LP evaluated at $\boldsymbol{x}$.

LEMMA 3. *There exists a positive constant $D \geq 1$ such that for any $\boldsymbol{x} \in \mathcal{X}$,*

$$D\left(\sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} - u^*\right) \geq d\left(\boldsymbol{x}, \mathcal{X}^*\right).$$

Note that $D \geq 1$ is necessary, since for every $\boldsymbol{x} \in \mathcal{X}$, $d\left(\boldsymbol{x}, \mathcal{X}^*\right) \geq \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} - u^*$.

PROOF. See Appendix A. □

2.3 Greedy with sublinear Safety Stocks (GSS)

Now we introduce the service policy, *Greedy with sublinear Safety Stocks (GSS)*, along with a variant, which we will prove to be asymptotically optimal.

2.3.1 GSS Policy in a Closed System

GSS. Let $p \in (\frac{1}{2}, 1)$. For a given r, define a weight function $w^r : \mathbb{R}_+ \to \mathbb{R}_+$ to be $w^r(X) = 1 \wedge \frac{X}{r^p}$. Let $\mathcal{M}$ denote the set of all pairs $(\boldsymbol{k}, i) \in \mathcal{K} \times \mathcal{I}$ such that $\boldsymbol{k} \in \mathcal{K}$ and $\boldsymbol{k} - \boldsymbol{e}_i \in \bar{\mathcal{K}}$. Given $\boldsymbol{X} = \{X_{\boldsymbol{k}'}, \boldsymbol{k}' \in \mathcal{K}\}$ and $(\boldsymbol{k}, i) \in \mathcal{M}$, define $\Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}) = w^r\left(X_{\boldsymbol{k}}\right) - w^r(X_{\boldsymbol{k}-\boldsymbol{e}_i})$. Under *GSS*, a customer of type i is placed into a server with configuration $\boldsymbol{k} - \boldsymbol{e}_i$ where $X_{\boldsymbol{k}-\boldsymbol{e}_i} > 0$ or $\boldsymbol{k} - \boldsymbol{e}_i = \boldsymbol{0}$, such that $\Delta_{(\boldsymbol{k},i)}(\boldsymbol{X})$ is minimal. Ties are broken arbitrarily.

Note that the *GSS* policy makes decisions based only the current system state. The parameter r which it uses is nothing else but the total number of customers in the system, which is, of course, a function of the state, and which happens to be constant in the closed system.

We now provide an intuitive explanation of the policy. Let f^r be the anti-derivative of w^r, so that

$$f^r(X) = \begin{cases} \frac{X^2}{2r^p}, & \text{if } X \in [0, r^p]; \\ X - \frac{r^p}{2}, & \text{if } X > r^p. \end{cases}$$

Let $F^r(\boldsymbol{X}) = \sum_{\boldsymbol{k}\in\mathcal{K}} f^r(X_{\boldsymbol{k}})$. Then w^r and $\Delta^r_{(\boldsymbol{k},i)}$ capture the first-order change in F^r. Suppose that the current system state is $\boldsymbol{X} = (X_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{K}}$. Then, placing a type-$i$ customer into a server with configuration $\boldsymbol{k} - \boldsymbol{e}_i$ only changes $X_{\boldsymbol{k}-\boldsymbol{e}_i}$ and $X_{\boldsymbol{k}}$: $X_{\boldsymbol{k}-\boldsymbol{e}_i}$ decreases by 1 (if $X_{\boldsymbol{k}-\boldsymbol{e}_i} > 0$), and $X_{\boldsymbol{k}}$ increases by 1. Thus, the first-order change in F^r is

$$\frac{d}{dX} f^r(X)\Big|_{X=X_{\boldsymbol{k}}} - \frac{d}{dX} f^r(X)\Big|_{X=X_{\boldsymbol{k}-\boldsymbol{e}_i}} = \Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}).$$

In this sense, *GSS* decreases F^r greedily, by placing a customer into a server that results in the largest (first-order) decrease in F^r.

The next lemma states that $F^r(\boldsymbol{X})$ only differs from $\sum_{\boldsymbol{k}} X_{\boldsymbol{k}}$ by $O(r^p)$. The proof is straightforward and omitted.

LEMMA 4. *For any $\boldsymbol{X} \in \mathbb{R}_+^{|\mathcal{K}|}$,*

$$\sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}} - \frac{|\mathcal{K}| r^p}{2} \leq F^r(\boldsymbol{X}) \leq \sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}.$$

Under the fluid scaling described earlier, the difference $O(r^p)$ between $F^r(\boldsymbol{X})$ and $\sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}$ becomes negligible, as it is of order $o(r)$. Thus, for a fluid-scaled process, minimizing $F^r(\boldsymbol{X})$ (what *GSS* tries to do) is "equivalent" to minimizing $\sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}$, when r is large.

2.3.2 GSS Policy in an Open System

First, we describe the "pure" *GSS* policy.

GSS. Let $p \in (\frac{1}{2}, 1)$. For a given system state $\boldsymbol{X}$, let $Z = Z(\boldsymbol{X})$ denote the total number of customers in the system. For a system with parameter r, define a weight function $\bar{w}^r(X) = \bar{w}^r(X; Z)$ as follows: $\bar{w}^r(X) = 1 \wedge \frac{X}{Z^p}$. (Note that $\bar{w}^r(X)$ generalizes the corresponding weight function $w^r(X) = 1 \wedge \frac{X}{r^p}$ for the closed system, because in the closed system with parameter r the total number of customers is constant $Z \equiv r$.) Let $\mathcal{M}$ denote the set of all pairs $(\boldsymbol{k}, i) \in \mathcal{K} \times \mathcal{I}$ such that $\boldsymbol{k} \in \mathcal{K}$ and $\boldsymbol{k} - \boldsymbol{e}_i \in \bar{\mathcal{K}}$. Given $\boldsymbol{X} = \{X_{\boldsymbol{k}'}, \boldsymbol{k}' \in \mathcal{K}\}$ and $(\boldsymbol{k}, i) \in \mathcal{M}$, define $\bar{\Delta}^r_{(\boldsymbol{k},i)}(\boldsymbol{X}) = \bar{w}^r\left(X_{\boldsymbol{k}}\right) - \bar{w}^r(X_{\boldsymbol{k}-\boldsymbol{e}_i})$. Under *GSS*, an arriving customer of type i is placed into a server with configuration $\boldsymbol{k} - \boldsymbol{e}_i$ where $X_{\boldsymbol{k}-\boldsymbol{e}_i} > 0$ or $\boldsymbol{k} - \boldsymbol{e}_i = \boldsymbol{0}$, such that $\bar{\Delta}_{(\boldsymbol{k},i)}(\boldsymbol{X})$ is minimal. Ties are broken arbitrarily.

In this paper, for the open system, we will analyze not the "pure" *GSS* policy, described above, but its slight modification, called *Modified GSS* (*GSS-M*).

GSS-M. Under this policy, a *token* of type i is generated immediately upon each service completion of type i, and is placed for "service" immediately according to *GSS*. The system state $\boldsymbol{X} = \{X_{\boldsymbol{k}}, \boldsymbol{k} \in \mathcal{K}\}$ account for both tokens of type i as well as actual type-i customers for all $i \in \mathcal{I}$. Each arriving type i customer first seeks to replace an existing token of type i already in "service" (chosen arbitrarily), and if there is none, it is placed for service according to *GSS*. Each token that is not replaced by an actual arriving customer before an independent exponentially distributed timeout with mean $1/\mu_0$, leaves the system. (This modification is the same as the one introduced in [14] for the *Greedy* algorithm, to obtain the *Greedy-M* policy.)

We emphasize that *GSS* and *GSS-M* do *not* require the knowledge of parameter r.

Since the system evolution under the *GSS-M* involves both actual customers and tokens, we need to define the Markov chain describing this evolution more precisely. A *complete server configuration* is defined (in the same way as in [14]) as a pair $(\boldsymbol{k}, \hat{\boldsymbol{k}})$, where vector $\boldsymbol{k} = (k_1, \ldots, k_I) \in \mathcal{K}$ gives the numbers of all customers (both actual and tokens) in a server, while vector $\hat{\boldsymbol{k}} \leq \boldsymbol{k}$, $\boldsymbol{k} \in \bar{\mathcal{K}}$, gives the numbers of actual customers only. The Markov process state at time t is the vector $\{X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}(t)\}$, where the index $(\boldsymbol{k}, \hat{\boldsymbol{k}})$ takes values that are all possible complete server configurations, and superscript r, as usual, indicates the system with parameter r. Note that $\boldsymbol{X}^r(t) = \{X^r_{\boldsymbol{k}}(t), \boldsymbol{k} \in \mathcal{K}\}$ can be considered as a "projection" of $\{X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}(t)\}$, with $X^r_{\boldsymbol{k}}(t) = \sum_{\hat{\boldsymbol{k}}:\hat{\boldsymbol{k}}\leq\boldsymbol{k}} X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}$ for each $\boldsymbol{k} \in \mathcal{K}$. Let $\hat{Y}^r_i(t)$, $\tilde{Y}^r_i(t)$, and $Y^r_i(t) = \hat{Y}^r_i(t) + \tilde{Y}^r_i(t)$ denote the total number of actual type-i customers, the total number of type-i tokens, and the total number of all (both actual and tokens) type-i customers in the rth system, respectively. The total number of actual customers of all types is then $Z^r(t) = \sum_i \hat{Y}^r_i(t)$. The behaviors of the processes $\{(Y^r_i(t), \hat{Y}^r_i(t)),\ t \geq 0\}$, are independent across all i, with $\hat{Y}^r_i(\infty)$ having Poisson distribution with mean $\rho_i r$. The following fact has the same proof as Lemma 11 in [14].

LEMMA 5. *The Markov chain $\{X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}(t)\},\ t \geq 0$, is irreducible and positive recurrent for each r.*

Remark. Informally, the reason (which is the same as in [14]) for considering a modified version of *GSS* instead of pure *GSS* in an open system is as follows. Recall that in a closed system, a customer migration can be also thought

of as its departure followed immediately by an arrival of the same type. As such, departures and arrivals in a closed system are perfectly "synchronized", which in particular means that in a closed system, for every departing customer, we always have the option of putting it right back into the server which it has just departed from. This means that a greedy control, pursuing minimization of a given objective function, cannot possibly increase (up to a first-order approximation) the objective function at every customer migration. In contrast, in an open system, departures and arrivals are not synchronized. Therefore, it is not immediately clear that a greedy algorithm will necessarily improve the objective. The tokens are introduced so that, informally speaking, the decisions on placements of new type-i arrivals are made somewhat "in advance", at the times of prior type-i departures. In this sense, the behavior of an open system "emulates" that of a corresponding closed system.

2.4 Main Results

THEOREM 6. *Let $p \in (\frac{1}{2}, 1)$. For each r, consider the closed system operating under* GSS *policy, in steady state. Then there exists some constant $C > 0$, not depending on r, such that*

$$\mathbb{P}\left(d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \le Cr^{p-1}\right) \to 1$$

as $r \to \infty$. Consequently, we have fluid-scale asymptotic optimality:

$$d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \implies 0.$$

THEOREM 7. *Let $p \in (\frac{1}{2}, 1)$. For each r, consider the open system operating under* GSS-M *policy, in steady state. Then there exists some constant $C > 0$, not depending on r, such that as $r \to \infty$,*

$$\mathbb{P}\left(d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \le Cr^{p-1}\right) \to 1, \tag{6}$$

and

$$r^{-p} \sum_i \tilde{Y}_i^r(\infty) \implies 0. \tag{7}$$

Consequently, we have fluid-scale asymptotic optimality:

$$d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \implies 0 \quad \text{and} \quad r^{-1} \sum_i \tilde{Y}_i^r(\infty) \implies 0.$$

3. CLOSED SYSTEM: ASYMPTOTIC OPTIMALITY OF GSS

We restrict our attention to closed systems and prove Theorem 6 in this section. As mentioned earlier, it is not sufficient to consider only the system states at the fluid scale, defined in Section 2.2. We also need the concept of *local fluid scaling*, introduced below. Proposition 9 – a key step in the proof of Theorem 6 – is established in Section 3.2. In Section 3.3, we construct an appropriate probability space, quantify the drift of F^r under *GSS* (cf. Propositions 14 and 15), and prove Theorem 6.

3.1 Local Fluid Scaling

Besides the fluid-scaled processes $\boldsymbol{x}^r(t)$ defined in Section 2.2, it is also convenient to consider the system dynamics at the *local fluid scale*. More precisely, for each r and t, define the corresponding *local fluid scale* process $\widetilde{\boldsymbol{x}}^r(t)$ by

$$\widetilde{\boldsymbol{x}}^r(t) = \frac{1}{r^p} \boldsymbol{X}^r(t).$$

In the asymptotic regime $r \to \infty$, recall that the fluid scale process $\boldsymbol{x}^r(\cdot)$ always lives in the compact set $\mathcal{X}$ (defined in Section 2.2). This is no longer true for the local fluid scale processes $\widetilde{\boldsymbol{x}}^r(\cdot)$: for a fixed t, $\{\widetilde{\boldsymbol{x}}^r(t)\}_r$ can be unbounded. However, at the local fluid scale, we will always consider the following weight function $\widetilde{w}$, which remains bounded.

Define the local-fluid-scale weight function $\widetilde{w} : \mathbb{R} \cup \{\infty\} \to \mathbb{R}_+$ to be $\widetilde{w}(\widetilde{x}) = 1 \wedge \widetilde{x}$. By convention, $1 < \infty$, so $\widetilde{w}$ is well-defined. Note that for every r, $\widetilde{w}(\widetilde{x}^r) = w^r(X^r)$, where $\widetilde{x}^r = X^r/r^p$. For $(\boldsymbol{k}, i) \in \mathcal{M}$, we can also define the weight difference at the local fluid scale to be

$$\Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}) = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}}) - \widetilde{w}(\widetilde{x}_{\boldsymbol{k}-\boldsymbol{e}_i}).$$

Remark. In the sequel, we will always use lower case x (or $\boldsymbol{x}$) to denote quantities at the fluid scale, $\widetilde{x}$ (or $\widetilde{\boldsymbol{x}}$) to denote quantities at the local fluid scale, and upper case X (or $\boldsymbol{X}$) to denote quantities without scaling.

3.2 Key Proposition

For a vector $\widetilde{\boldsymbol{x}} \in (\mathbb{R}_+ \cup \{\infty\})^{|\mathcal{K}|}$ with components being possibly infinite, we can define the concept of a *Strictly Improving (SI) pair associated with $\widetilde{\boldsymbol{x}}$.*

DEFINITION 8 (STRICTLY IMPROVING (SI) PAIR). *For $(\boldsymbol{k}, i)$, $(\boldsymbol{k}', i) \in \mathcal{M}$, $\{(\boldsymbol{k}, i), (\boldsymbol{k}', i)\}$ is an SI pair associated with $\widetilde{\boldsymbol{x}}$ if*

(a) $k_i \ge 1$, $\widetilde{x}_{\boldsymbol{k}} > 0$;

(b) either $\boldsymbol{k}' = \boldsymbol{e}_i$, or $[k'_i > 0$ and $\widetilde{x}_{\boldsymbol{k}'-\boldsymbol{e}_i} > 0]$; and

(c) $\Delta_{(\boldsymbol{k}',i)} < \Delta_{(\boldsymbol{k},i)}$.

The idea of SI pairs is as follows. Suppose that the current system state is $\boldsymbol{X}^r$, and a type-i customer just completed its service requirement at a server with configuration $\boldsymbol{k}$. Then the first-order change in F^r is $-\Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}^r)$. Suppose that this customer is then placed into a server with configuration $\boldsymbol{k}'$, under *GSS*. Then, the total (first-order) change in F^r after this transition is $\Delta^r_{(\boldsymbol{k}',i)}(\boldsymbol{X}^r) - \Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}^r)$, or $\Delta_{(\boldsymbol{k}',i)}(\widetilde{\boldsymbol{x}}^r) - \Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}^r)$. The existence of an SI pair ensures that we can always improve (up to first order) the current value of F^r.

Recall that for any feasible system state $\boldsymbol{X}^r$, $\boldsymbol{x}^r = \boldsymbol{X}^r/r$ denotes the fluid-scale system state, and $\widetilde{\boldsymbol{x}}^r = \boldsymbol{X}^r/r^p$ denotes the associated state at the local fluid scale. The following proposition establishes that whenever $\boldsymbol{x}^r$ is sufficiently far away from optimality, an SI pair exists.

PROPOSITION 9. *Let $D > 0$ be the same as in Lemma 3. Then, there exist a positive constant ε such that the following holds. For sufficiently large r, if $d(\boldsymbol{x}^r, \mathcal{X}^*) \ge 2D|\mathcal{K}|r^{p-1}$, then there exists an SI pair $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ (possibly depending on r) associated with $\widetilde{\boldsymbol{x}}^r = (\widetilde{x}^r_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{K}}$, and furthermore, $\widetilde{x}^r_{\boldsymbol{k}} \ge \varepsilon$, $\widetilde{x}^r_{\boldsymbol{k}'-\boldsymbol{e}_i} \ge \varepsilon$, and $\Delta_{(\boldsymbol{k}',i)}(\widetilde{\boldsymbol{x}}^r) - \Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}^r) \le -\varepsilon$.*

Proposition 9 follows from the two lemmas below.

LEMMA 10. *Consider any sequence $\{\boldsymbol{x}^r\}$ and the associated states $\widetilde{\boldsymbol{x}}^r$. Let $\boldsymbol{x} \in \mathcal{X}$ be a limit point of the sequence $\{\boldsymbol{x}^r\}$, so that the the subsequence $\{r_n\}$ of $\{r\}$ satisfies $\boldsymbol{x}^{r_n} \to \boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}^{r_n} \to \widetilde{\boldsymbol{x}}$ as $n \to \infty$, with some components of $\widetilde{\boldsymbol{x}}$ being possibly infinite. If there is no SI pair associated with $\widetilde{\boldsymbol{x}}$, then $\boldsymbol{x} \in \mathcal{X}^*$, i.e. $\boldsymbol{x}$ is an optimal solution of LP.*

PROOF OF LEMMA 10. Suppose that there is no SI pair associated with $\widetilde{\boldsymbol{x}}$. We will show that $\boldsymbol{x} \in \mathcal{X}^*$, i.e., $\boldsymbol{x}$ is an optimal solution of the linear program LP. To this end, we will use Lemma 2. In particular, we will construct $\eta_i \geq 0$, $i \in \mathcal{I}$ such that

(i) $\sum_{i\in\mathcal{I}} k_i\eta_i \leq 1$ for all $\boldsymbol{k} \in \mathcal{K}$, and

(ii) if $\sum_{i\in\mathcal{I}} k_i\eta_i < 1$, then $\widetilde{x}_{\boldsymbol{k}} < 1$.

Note that condition (ii) here is stronger than condition (ii) in Lemma 2.

Let $\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{e}_i})$ for all $i \in \mathcal{I}$. Then clearly $\eta_i \in [0,1]$ for all $i \in \mathcal{I}$. We first show that condition (i) holds. To this end, we prove the following stronger statement: if $\boldsymbol{k} \in \mathcal{K}$ is such that $k_i \geq 1$ implies $\eta_i > 0$, then $\sum_{i\in\mathcal{I}} k_i\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}})$. Suppose not. Let $\boldsymbol{k} \in \mathcal{K}$ be a minimal counterexample, so that

$$\sum_{i\in\mathcal{I}} k_i\eta_i \neq \widetilde{w}(\widetilde{x}_{\boldsymbol{k}}), \tag{8}$$

and for each $i \in \mathcal{I}$, $k_i \geq 1$ implies $\eta_i > 0$. Note that $\sum_{i\in\mathcal{I}} k_i \geq 2$, since $\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{e}_i})$ for each $i \in \mathcal{I}$, by definition. Thus, there exists $i \in \mathcal{I}$ such that $\eta_i > 0$, $\boldsymbol{k}' = \boldsymbol{k} - \boldsymbol{e}_i \in \mathcal{K}$, and

$$\sum_{i\in\mathcal{I}} k'_i\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}'}). \tag{9}$$

Subtracting Eq. (9) from Eq. (8), we get that

$$\Delta_{(\boldsymbol{k},i)} = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}}) - \widetilde{w}(\widetilde{x}_{\boldsymbol{k}'}) \neq \eta_i.$$

Thus either $\Delta_{(\boldsymbol{k},i)} > \eta_i$, or $\Delta_{(\boldsymbol{k},i)} < \eta_i$. If $\Delta_{(\boldsymbol{k},i)} > \eta_i$, we verify that $\{(\boldsymbol{k},i),(\boldsymbol{e}_i,i)\}$ is an SI pair associated with $\widetilde{\boldsymbol{x}}$. First, conditions (b) and (c) in Definition 8 are automatically satisfied. Second, $\Delta_{(\boldsymbol{k},i)} > \eta_i > 0$. In particular, $\widetilde{x}_{\boldsymbol{k}} > 0$. We also have $k_i \geq 1$, so condition (a) in Definition 8 is also satisfied.

If $\Delta_{(\boldsymbol{k},i)} < \eta_i$, we verify that $\{(\boldsymbol{e}_i,i),(\boldsymbol{k},i)\}$ is an SI pair associated with $\widetilde{\boldsymbol{x}}$. First, condition (c) in Definition 8 is automatically satisfied. Second, since $\eta_i > 0$, $\widetilde{x}_{\boldsymbol{e}_i} > 0$. Thus condition (a) in Definition 8 is satisfied. Finally, $k_i \geq 1$ by assumption, so to verify condition (b), we only need to verify that $\widetilde{x}_{\boldsymbol{k}-\boldsymbol{e}_i} > 0$. Since $\sum_{i\in\mathcal{I}} k_i \geq 2$, $\sum_{i\in\mathcal{I}} k'_i \geq 1$. This implies that there exists $i' \in \mathcal{I}$ such that $k'_{i'} \geq 1$. Thus $k_{i'} \geq k'_{i'} \geq 1$, so $\eta_{i'} > 0$. By Eq. (9), $\widetilde{w}(\widetilde{x}_{\boldsymbol{k}'}) \geq \eta_{i'} > 0$, so $\widetilde{x}_{\boldsymbol{k}'} > 0$. Thus, condition (b) in Definition 8 is verified.

In either case, we have an SI pair associated with $\widetilde{\boldsymbol{x}}$, contradicting the assumption that there is no SI pair associated with $\widetilde{\boldsymbol{x}}$. Thus, for all $\boldsymbol{k} \in \mathcal{K}$ such that $k_i \geq 1$ implies $\eta_i > 0$,

$$\sum_{i\in\mathcal{I}} k_i\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}}).$$

For all $\boldsymbol{k} \in \mathcal{K}$, we can find $\boldsymbol{k}' \leq \boldsymbol{k}$ such that $\boldsymbol{k}' \in \mathcal{K}$, $k'_i \geq 1$ implies $\eta_i > 0$, and $\sum_{i\in\mathcal{I}} k_i\eta_i = \sum_{i\in\mathcal{I}} k'_i\eta_i$. Thus,

$$\sum_{i\in\mathcal{I}} k_i\eta_i = \sum_{i\in\mathcal{I}} k'_i\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}'}) \leq 1.$$

This establishes condition (i).

We now establish condition (ii). Suppose that condition (ii) does not hold. Let $\boldsymbol{k} \in \mathcal{K}$ be minimal such that

$$\widetilde{x}_{\boldsymbol{k}} \geq 1, \quad \text{and} \quad \sum_{i\in\mathcal{I}} k_i\eta_i < 1.$$

First, note that $\boldsymbol{k} \neq \boldsymbol{e}_i$ for any $i \in \mathcal{I}$, because if $\eta_i < 1$, then

$$1 > \eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{e}_i}) = 1 \wedge \widetilde{x}_{\boldsymbol{e}_i}.$$

Thus $\sum_{i\in\mathcal{I}} k_i \geq 2$. Second, if $\eta_i > 0$ for all $i \in \mathcal{I}$ with $k_i \geq 1$, then from the proof of condition (i), we have that

$$1 > \sum_{i\in\mathcal{I}} k_i\eta_i = \widetilde{w}(\widetilde{x}_{\boldsymbol{k}}) = 1 \wedge \widetilde{x}_{\boldsymbol{k}},$$

so we have $\widetilde{x}_{\boldsymbol{k}} < 1$, reaching a contradiction. Thus, there exists $i \in \mathcal{I}$ such that $\eta_i = 0$ and $k_i \geq 1$. Let $\boldsymbol{k}' = \boldsymbol{k} - \boldsymbol{e}_i$. Then $\boldsymbol{k}' \in \mathcal{K}$, since

$$\sum_{i\in\mathcal{I}} k'_i = \sum_{i\in\mathcal{I}} k_i - 1 \geq 1.$$

Since $\eta_i = 0$,

$$\sum_{i\in\mathcal{I}} k'_i\eta_i = \sum_{i\in\mathcal{I}} k_i\eta_i < 1.$$

By minimality of $\boldsymbol{k}$, we must have $\widetilde{x}_{\boldsymbol{k}'} < 1$. Thus, $\widetilde{w}(\widetilde{x}_{\boldsymbol{k}'}) = 1 \wedge \widetilde{x}_{\boldsymbol{k}'} < 1$, and $\widetilde{w}(\widetilde{x}_{\boldsymbol{k}}) = 1 \wedge \widetilde{x}_{\boldsymbol{k}} = 1$. This implies that

$$\Delta_{(\boldsymbol{k},i)} > 0 = \eta_i,$$

and that $\{(\boldsymbol{k},i),(\boldsymbol{e}_i,i)\}$ is an SI pair associated with $\widetilde{\boldsymbol{x}}$. This is a contradiction, so condition (ii) is established. □

LEMMA 11. *Consider any sequence $\{\boldsymbol{x}^r\}$ and associated states $\widetilde{\boldsymbol{x}}^r$. Let $\boldsymbol{x}^{r_n}$, $\boldsymbol{x}$, $\widetilde{\boldsymbol{x}}^{r_n}$ and $\widetilde{\boldsymbol{x}}$ be the same as in Lemma 10. If for all sufficiently large n, $d(\boldsymbol{x}^{r_n}, \mathcal{X}^*) \geq 2D|\mathcal{K}|r_n^{p-1}$, then there is an SI pair associated with $\widetilde{\boldsymbol{x}}$.*

PROOF OF LEMMA 11. We prove the lemma by contradiction. Suppose that the lemma is not true, then for sufficiently large n, $d(\boldsymbol{x}^{r_n}, \mathcal{X}^*) \geq 2D|\mathcal{K}|r_n^{p-1}$, and there is no SI pair associated with $\widetilde{\boldsymbol{x}}$. By Lemma 10, $\boldsymbol{x}$ is an optimal solution of LP, and from the proof of Lemma 10, $\boldsymbol{\eta} = (\eta_i)_{i\in\mathcal{I}}$ is an optimal dual solution of LP, where $\eta_i = \widetilde{x}_{\boldsymbol{e}_i}$ for all $i \in \mathcal{I}$.

For a given r, consider the following linear program, which we call LP^r.

$$\text{Minimize} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}_{\boldsymbol{k}} \tag{10}$$

$$\text{subject to} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} k_i\widetilde{x}_{\boldsymbol{k}} = \rho_i r^{1-p}, \quad \text{for all } i \in \mathcal{I}, \tag{11}$$

$$\widetilde{x}_{\boldsymbol{k}} \geq 0, \qquad \text{for all } \boldsymbol{k} \in \mathcal{K}. \tag{12}$$

LP^r is just a scaled version of LP, defined in Section 2.2. For each r, the feasible set of LP^r is $r^{1-p}\mathcal{X}$, its set of optimal solutions is $r^{1-p}\mathcal{X}^*$, and its optimal value is $r^{1-p}u^*$. $r^{1-p}\boldsymbol{x}$ is an optimal solution of LP^r, and $\boldsymbol{\eta}$ is an optimal dual solution. Furthermore, by Lemma 3, for sufficiently large n,

$$\begin{aligned}\sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}^{r_n}_{\boldsymbol{k}} - r^{1-p}u^* &= r^{1-p}\left(\sum_{\boldsymbol{k}\in\mathcal{K}} x^{r_n}_{\boldsymbol{k}} - u^*\right)\\ &\geq r^{1-p}d(\boldsymbol{x}^{r_n}, \mathcal{X}^*)/D\\ &\geq r^{1-p}\cdot(2D|\mathcal{K}|r^{p-1})/D \geq 2|\mathcal{K}|.\end{aligned}$$

For each n, consider the Lagrangian $L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta})$ of LP^{r_n}, evaluated at $\widetilde{\boldsymbol{x}}^{r_n}$ and $\boldsymbol{\eta}$:

$$L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta}) = \sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}^{r_n}_{\boldsymbol{k}} + \sum_{i\in\mathcal{I}} \eta_i\left(\rho_i r_n^{1-p} - \sum_{\boldsymbol{k}\in\mathcal{K}} k_i\widetilde{x}^{r_n}_{\boldsymbol{k}}\right).$$

We calculate the Lagrangian in two ways. First, by feasibility of $\widetilde{\boldsymbol{x}}^{r_n}$, $L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta}) = \sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}^{r_n}_{\boldsymbol{k}}$. Second, we rewrite $L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta})$ as

$$L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta}) = r_n^{1-p} \sum_{i\in\mathcal{I}} \rho_i \eta_i + \sum_{\boldsymbol{k}\in\mathcal{K}} \left(1 - \sum_{i\in\mathcal{I}} k_i \eta_i\right) \widetilde{x}^{r_n}_{\boldsymbol{k}}.$$

The first term on the RHS equals $r_n^{1-p} u^*$, by the dual optimality of $\boldsymbol{\eta}$. For the second term on the RHS, note that in the proof of Lemma 10, we have established that for all $\boldsymbol{k} \in \mathcal{K}$, $\sum_{i\in\mathcal{I}} k_i \eta_i \le 1$, and if $\sum_{i\in\mathcal{I}} k_i \eta_i < 1$, then $\widetilde{x}_{\boldsymbol{k}} < 1$. Since $\widetilde{\boldsymbol{x}}^{r_n} \to \widetilde{\boldsymbol{x}}$, for all sufficiently large n, if $\sum_{i\in\mathcal{I}} k_i \eta_i < 1$, then $\widetilde{x}^{r_n}_{\boldsymbol{k}} \le 1$. Thus for all sufficiently large n,

$$\sum_{\boldsymbol{k}\in\mathcal{K}} \left(1 - \sum_{i\in\mathcal{I}} k_i \eta_i\right) \widetilde{x}^{r_n}_{\boldsymbol{k}} \le |\mathcal{K}|,$$

and

$$\sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}^{r_n}_{\boldsymbol{k}} = L(\widetilde{\boldsymbol{x}}^{r_n}, \boldsymbol{\eta}) \le r_n^{1-p} u^* + |\mathcal{K}|,$$

contradicting the fact that

$$\sum_{\boldsymbol{k}\in\mathcal{K}} \widetilde{x}^{r_n}_{\boldsymbol{k}} - r_n^{1-p} u^* \ge 2|\mathcal{K}|$$

for sufficiently large n. This establishes Lemma 11. □

Proof of Proposition 9. We are now ready to prove Proposition 9. Suppose that the proposition does not hold. Then for all $\varepsilon > 0$, there exist infinitely many r and $\boldsymbol{x}^r$ such that $d(\boldsymbol{x}^r, \mathcal{X}^*) \ge 2D|\mathcal{K}|r^{p-1}$, and for all SI pairs (if any) $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ of $\widetilde{\boldsymbol{x}}^r$, either $\widetilde{x}^r_{\boldsymbol{k}} < \varepsilon$, or $\widetilde{x}^r_{\boldsymbol{k}'-\boldsymbol{e}_i} < \varepsilon$, or $\Delta_{(\boldsymbol{k}',i)}(\widetilde{\boldsymbol{x}}^r) - \Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}^r) > -\varepsilon$. Thus, we can find a subsequence $\{r_n\}$ of $\{r\}$ and states $\boldsymbol{x}^{r_n}$ such that

1. $\boldsymbol{x}^{r_n} \to \boldsymbol{x} \in \mathcal{X}$ as $n \to \infty$,
2. $\widetilde{\boldsymbol{x}}^{r_n} \to \widetilde{\boldsymbol{x}}$ as $n \to \infty$, with some components of $\widetilde{\boldsymbol{x}}$ being possibly infinite,
3. $d(\boldsymbol{x}^{r_n}, \mathcal{X}^*) \ge 2D|\mathcal{K}|r_n^{p-1}$ for all n, and
4. for all SI pairs $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ associated with $\widetilde{\boldsymbol{x}}^{r_n}$ (if any), either $\widetilde{x}^{r_n}_{\boldsymbol{k}} < 1/n$, or $\widetilde{x}^{r_n}_{\boldsymbol{k}'-\boldsymbol{e}_i} < 1/n$, or $\Delta_{(\boldsymbol{k}',i)}(\widetilde{\boldsymbol{x}}^{r_n}) - \Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}^{r_n}) > -1/n$.

From Property 4, we can deduce that $\widetilde{\boldsymbol{x}}$ does not have an SI pair. But by Property 3, this contradicts Lemma 11. This establishes Proposition 9. □

3.3 Proof of Theorem 6

We will assume WLOG the following construction of the probability space. For each $(\boldsymbol{k}, i) \in \mathcal{M}$, consider an independent unit-rate Poisson process $\{\Pi_{(\boldsymbol{k},i)}(t),\ t \ge 0\}$. Assume that, for each r, the Markov process $\boldsymbol{X}^r(\cdot)$ is driven by this common set of Poisson processes $\Pi_{(\boldsymbol{k},i)}(\cdot)$, as follows. For each $(\boldsymbol{k}, i) \in \mathcal{M}$, let us denote by $D^r_{(\boldsymbol{k},i)}(t)$ the total number of type-i service completions from servers of configuration $\boldsymbol{k}$, in the time interval $[0, t]$. Then

$$D^r_{(\boldsymbol{k},i)}(t) = \Pi_{(\boldsymbol{k},i)}\left(\int_0^t X^r_{\boldsymbol{k}}(\xi) k_i \mu_i d\xi\right). \tag{13}$$

Lemma 12. *Let $T > 0$ be fixed. With probability 1, the following property holds. Consider any sequence $\{t_0^r\}_r$ with $t_0^r \in [0, Tr^{2-p}]$. Then for any $\xi \in [0, 1]$, and for any $(\boldsymbol{k}, i) \in \mathcal{M}$,*

$$\frac{1}{r^{2p-1}}\left(\Pi_{(\boldsymbol{k},i)}\left(t_0^r + \xi r^{2p-1}\right) - \Pi_{(\boldsymbol{k},i)}\left(t_0^r\right)\right) \to \xi$$

as $r \to \infty$. The convergence is uniform over t_0^r, ξ, and $(\boldsymbol{k}, i)$ in the following sense. For any $\varepsilon > 0$, there exists $r(\varepsilon)$ such that for all $r \ge r(\varepsilon)$, $\xi \in [0, 1]$, $(\boldsymbol{k}, i) \in \mathcal{M}$, and $t_0^r \in [0, Tr^{2-p}]$,

$$\max_{(\boldsymbol{k},i),\xi,t_0^r} \left|\frac{1}{r^{2p-1}}\left(\Pi_{(\boldsymbol{k},i)}\left(t_0^r + \xi r^{2p-1}\right) - \Pi_{(\boldsymbol{k},i)}\left(t_0^r\right)\right) - \xi\right| < \varepsilon.$$

The proof of Lemma 12 depends on simple large-deviation type estimates for Poisson random variables. The idea is essentially the same as that of Lemma 4.3 in [11]: we partition the interval $[0, Tr^{2p-1}]$ into subintervals of length $r^{p-1/2}$, and for each of them write the probability that the average increase rate of $\Pi_{(\boldsymbol{k},i)}$ lies outside $(1-\varepsilon, 1+\varepsilon)$. These probabilities are $\exp(-\text{poly}(r))$, and we only have $\text{poly}(r)$ such subintervals (here $\text{poly}(r)$ means a polynomial in r). This is true for any $\varepsilon > 0$. We can then cover *any* subinterval of length r^{2p-1} by these subintervals of length $r^{p-1/2}$. We omit a detailed proof here.

The following corollary is a simple consequence of Lemma 12.

Corollary 13. *Let T be fixed. With probability 1, the following holds. For sufficiently large r,*

$$\max_{\substack{\xi\in[0,1],\\ t_0^r\in[0,Tr^{1-p}]}} d\left(\boldsymbol{X}^r(t_0^r + \xi r^{p-1}), \boldsymbol{X}^r(t_0^r)\right) \le 2\bar{\mu}|\mathcal{K}|r^p, \tag{14}$$

where $\bar{\mu} = \max_{i\in\mathcal{I}} \mu_i$, and μ_i is the service rate for type-i customers.

Proof. Consider the probability-1 event in Lemma 12, in which we can and do replace T with $2\bar{\mu}T$. (We do this because the total "instantaneous" rate of all transitions is upper bounded by $2\bar{\mu}r$.) The rate of departure of type-i customers is $\rho_i \mu_i r \le \rho_i \bar{\mu} r$, and the total rate of customer departure is no greater than $\sum_{i\in\mathcal{I}} \rho_i \bar{\mu} r = \bar{\mu} r$. Thus, for each $\boldsymbol{k} \in \mathcal{K}$, the rate of change in $X_{\boldsymbol{k}}$ is at most $\bar{\mu} r$. For an interval of length r^{p-1}, the total change in $X_{\boldsymbol{k}}$ is at most $O(r \cdot r^{p-1}) = O(r^p)$. More precisely, with probability 1, for each $\boldsymbol{k} \in \mathcal{K}$,

$$\limsup_{r\to\infty} \frac{1}{r^p} \max_{\substack{\xi\in[0,1],\\ t_0^r\in[0,Tr^{1-p}]}} \left|X^r_{\boldsymbol{k}}(t_0^r + \xi r^{p-1}) - X^r_{\boldsymbol{k}}(t_0^r)\right| \le \bar{\mu}.$$

Thus, for sufficiently large r, and for each $\boldsymbol{k} \in \mathcal{K}$,

$$\max_{\substack{\xi\in[0,1],\\ t_0^r\in[0,Tr^{1-p}]}} \left|X^r_{\boldsymbol{k}}(t_0^r + \xi r^{p-1}) - X^r_{\boldsymbol{k}}(t_0^r)\right| \le 2\bar{\mu} r^p.$$

Summing over the above expression establishes the corollary. □

Proposition 14. *There exist positive constants C_1 and δ such that the following holds. Let $T > 0$ be given. Then w.p.1, for all sufficiently large r, and for any interval $[t_0, t_0 + r^{p-1}] \subset [0, Tr^{1-p}]$, if $d(\boldsymbol{x}^r(t_0), \mathcal{X}^*) \ge C_1 r^{p-1}$, then*

$$F^r\left(\boldsymbol{X}^r(t_0 + r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(t_0)\right) \le -\delta r^{2p-1}.$$

PROOF. The proof idea is as follows. Consider the increase in F^r at each state transition. For concreteness, suppose that the current system state is $\boldsymbol{X}^r$, and a type-i customer just completed its service requirement on a server with configuration $\boldsymbol{k}$, and is placed into a server with configuration $\boldsymbol{k}'$. Then it is a simple calculation to see that the increase in F^r is at most

$$\Delta^r_{(\boldsymbol{k}',i)}(\boldsymbol{X}^r) - \Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}^r) + 4r^{-p}.$$

The term $\Delta^r_{(\boldsymbol{k}',i)}(\boldsymbol{X}^r) - \Delta^r_{(\boldsymbol{k},i)}(\boldsymbol{X}^r)$ captures the first-order increase in F^r, and the term $4r^{-p}$ bounds the second-order increase in F^r. We will see that over an interval of length r^{p-1}, the increase in F^r due to first-order terms is at most $-O(r^{2p-1})$, and the increase due to second-order terms is at most a constant. We now proceed to the formal proof.

From now on, we work with the probability-1 event defined in Lemma 12, under which

$$\frac{1}{r^{2p-1}}\left(\Pi_{(\boldsymbol{k},i)}\left(t_0+\xi r^{2p-1}\right) - \Pi_{(\boldsymbol{k},i)}\left(t_0\right)\right) \to \xi$$

as $r\to\infty$, uniformly over t_0, ξ, and $(\boldsymbol{k}, i)$. Let $C_1 = 2(\bar{\mu} + D)|\mathcal{K}|$, where $\bar{\mu} = \max_{i\in\mathcal{I}}\mu_i$ and D is the same as in Lemma 3. Let $\varepsilon > 0$ be the same as in Proposition 9, and let $\delta > 0$ be such that $\delta < \frac{1}{8}\mu_i\varepsilon^2$ for all $i \in \mathcal{I}$.

Claim that for all sufficiently large r, and for any interval $[t_0, t_0 + r^{p-1}] \subset [0, Tr^{1-p}]$, if $d\left(\boldsymbol{x}^r(t_0), \mathcal{X}^*\right) \geq C_1 r^{p-1}$, then

$$F^r\left(\boldsymbol{X}^r(t_0 + r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(t_0)\right) \leq -\delta r^{2p-1}.$$

Suppose the contrary. Then there exist a subsequence of $\{r\}$ (which, with an abuse of notation, we still index by r), along which we have some $[t_0^r, t_0^r + r^{p-1}] \subset [0, Tr^{1-p}]$, such that $d\left(\boldsymbol{x}^r(t_0^r), \mathcal{X}^*\right) \geq C_1 r^{p-1}$, and

$$F^r\left(\boldsymbol{X}^r(t_0^r + r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(t_0^r)\right) > -\delta r^{2p-1}. \tag{15}$$

First, for sufficiently large r, and for all $\xi \in [0,1]$, there exists a SI pair $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ associated with $\boldsymbol{x}^r(t_0^r + \xi r^{p-1})$ (possibly depending on r and ξ), such that

$$\widetilde{x}^r_{\boldsymbol{k}}(t_0^r + \xi r^{p-1}) \geq \varepsilon, \quad \widetilde{x}^r_{\boldsymbol{k}'-\boldsymbol{e}_i}(t_0^r + \xi r^{p-1}) \geq \varepsilon, \text{ and} \tag{16}$$

$$\Delta_{(\boldsymbol{k}',i)}(\widetilde{\boldsymbol{x}}^r(t_0^r + \xi r^{p-1})) - \Delta_{(\boldsymbol{k},i)}(\widetilde{\boldsymbol{x}}^r(t_0^r + \xi r^{p-1})) \leq -\varepsilon. \tag{17}$$

By Corollary 13, for all $\xi \in [0,1]$, $d\left(\boldsymbol{X}^r(t_0^r + \xi r^{p-1}), \boldsymbol{X}^r(t_0^r)\right) \leq 2\bar{\mu}|\mathcal{K}|r^p$. Using triangle inequality and choosing $C_1 > 2(\bar{\mu} + D)|\mathcal{K}|$, we have that for sufficiently large r, and for all $\xi \in [0,1]$,

$$d\left(\boldsymbol{x}^r(t_0^r + \xi r^{p-1}), \mathcal{X}^*\right) \geq 2D|\mathcal{K}|r^{p-1}.$$

(16) and (17) now follow from Proposition 9.

Fix a sufficiently large r so that (16) and (17) hold. We then consider the first-order change in F^r over the interval $[t_0^r, t_0^r + r^{p-1}]$ (i.e., the difference of Δ). To do this, we partition $[t_0^r, t_0^r + r^{p-1}]$ into subintervals of length $c\varepsilon r^{p-1}$, with $c > 0$ chosen small enough so that on each subinterval, there exists a *fixed* SI pair $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ such that (16) and (17) hold for this SI pair, and with ε replaced by $\varepsilon/2$. We now argue that this can be done. Consider the first such subinterval, for example. By Lemma 12, for sufficiently large r, the number of state transitions over this subinterval is at most $(c\varepsilon r^{p-1}) \cdot O(r) = O(\varepsilon r^p) < \frac{1}{8}\varepsilon r^p$, by choosing a sufficiently small c. This implies that for each $\boldsymbol{k} \in \mathcal{K}$, the change in $\tilde{x}^r_{\boldsymbol{k}}$ over this subinterval is at most $\frac{1}{8}\varepsilon$. Thus, (16) and (17) hold for an SI pair associated with $\tilde{\boldsymbol{x}}^r(t_0^r)$, with ε replaced by $\varepsilon/2$. The same argument holds for other subintervals.

Now concentrate on the subinterval $[t_0^r, t_0^r + c\varepsilon r^{p-1}]$, and a corresponding SI pair $\{(\boldsymbol{k}', i), (\boldsymbol{k}, i)\}$ associated with $\tilde{\boldsymbol{x}}^r(t_0^r)$ for which (16) and (17) hold on this subinterval with ε replaced by $\varepsilon/2$. The number of type-i departures from servers of configuration $\boldsymbol{k}$ is at least $\mu_i \cdot \frac{\varepsilon r^p}{2} \cdot (c\varepsilon r^{p-1}) = \frac{1}{2}c\mu_i\varepsilon^2 r^{2p-1}$. At each such departure, the first-order increase (due to the difference of Δ) in F^r is at most $-\varepsilon/2$, since *GSS* results in a smaller first-order increase than moving the departure to a server with configuration $\boldsymbol{k}' - \boldsymbol{e}_i$. Summing over all such increases over type-i departures gives a first-order increase in F^r which is at most

$$-\frac{\varepsilon}{2} \cdot \left(\frac{1}{2}c\mu_i\varepsilon^2 r^{2p-1}\right) \leq -2c\varepsilon\delta r^{2p-1}.$$

Exactly the same argument holds for other subintervals, so the total first-order increase in F^r is at most $-2\delta r^{2p-1}$.

Finally, consider the second-order increase in F^r. As discussed at the beginning of the proof, the second-order increase in F^r at each state transition is at most $4r^{-p}$. For sufficiently large r, the total number of state transitions over the interval $[t_0^r, t_0^r + r^{p-1}]$ is at most $r^{p-1} \cdot O(r) = O(r^p)$, and hence the total second-order increase in F^r is at most $(4r^{-p}) \cdot O(r^p) = O(1)$. Thus, for sufficiently large r,

$$F^r\left(\boldsymbol{X}^r(t_0^r + r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(t_0^r)\right) \leq -2\delta r^{2p-1} + O(1) \leq -\delta r^{2p-1}.$$

This contradicts (15), and we have established the proposition. □

PROPOSITION 15. *There exist positive constants C and T such that as $r\to\infty$,*

$$\mathbb{P}\left(d\left(\boldsymbol{x}^r(Tr^{1-p}), \mathcal{X}^*\right) \leq Cr^{p-1}\right) \to 1.$$

PROOF SKETCH. The proof is very intuitive. We keep track of the evolution of F^r on the interval $[0, Tr^{1-p}]$ subdivided into r^{p-1}-long subintervals. W.p.1., for all sufficiently large r, the following is true for each subinterval $[t_0, t_0 + r^{p-1}]$: F^r decreases by at least δr^{2p-1} if $d\left(\boldsymbol{x}^r(t_0), \mathcal{X}^*\right) \geq C_1 r^{p-1}$ (by Proposition 14), and it can never increase by more than $C_3 r^p$. Therefore, if we choose T large enough, then $d\left(\boldsymbol{x}^r(t), \mathcal{X}^*\right) < C_1 r^{p-1}$ at some time $t \in [0, Tr^{1-p}]$ (because otherwise F^r would become negative), and $d\left(\boldsymbol{x}^r(t), \mathcal{X}^*\right) = O(r^{p-1})$ thereafter. We refer the readers to Appendix B for details. □

Proof of Theorem 6. Theorem 6 is now a simple consequence of Proposition 15. For each r, consider $\boldsymbol{x}^r(\cdot)$ in the stationary regime. In particular, for any $T > 0$, $\boldsymbol{x}^r(Tr^{1-p})$ has the same distribution as $\boldsymbol{x}^r(\infty)$. Therefore, by Proposition 15,

$$\mathbb{P}\left(d\left(\boldsymbol{x}^r(\infty), \mathcal{X}^*\right) \leq Cr^{p-1}\right) \to 1,$$

as $r\to\infty$. This completes the proof of Theorem 6. □

4. OPEN SYSTEM: ASYMPTOTIC OPTIMALITY OF (MODIFIED) GSS

We prove Theorem 7 in this section. The proof "extends" that of Theorem 6. The main additional step is Theorem 18, which shows that in steady state, for each $i \in \mathcal{I}$, $\tilde{Y}_i^r(t)$ the number of tokens of type-i, remains $o(r^p)$ with high probability, over $O(r^{1-p})$-long intervals. As a starting point, we need the following facts.

THEOREM 16. *Consider the sequence (in r) of open systems in steady state. Consider any fixed i. There exists a positive constant c such that, uniformly on all r,*

$$\mathbb{E}\exp\{\|r^{-1/2}(\hat{Y}_i^r(\infty)-\rho_i r,\tilde{Y}_i^r(\infty))\|\}\le c.$$

PROOF. See Appendix C. □

For our purposes, the following corollary will suffice.

COROLLARY 17. *Consider the sequence (in r) of open systems in steady state. Consider any fixed i. Then, for any $q>1/2$,*

$$\|r^{-q}(\hat{Y}_i^r(\infty)-\rho_i r,\tilde{Y}_i^r(\infty))\|\Longrightarrow 0.$$

Next we show that the property of Corollary 17 holds not just at a given time, but uniformly on a $O(r^{1-q})$-long interval.

THEOREM 18. *Consider the sequence (in r) of open systems in stationary regime. Consider any fixed i. Let $q>1/2$ and $T>0$ be fixed. Then, as $r\to\infty$,*

$$\sup_{t\in[0,Tr^{1-q}]}\|r^{-q}(\hat{Y}_i^r(t)-\rho_i r,\tilde{Y}_i^r(t))\|\Longrightarrow 0, \tag{18}$$

and, consequently,

$$\sup_{t\in[0,Tr^{1-q}]} r^{-q}\|Z^r(t)-r\|\Longrightarrow 0. \tag{19}$$

Clearly, the statement of Theorem 18 is equivalent to the following one: *Any subsequence of $\{r\}$ contains a further subsequence along which w.p.1,*

$$\sup_{t\in[0,Tr^{1-q}]}\|r^{-q}(\hat{Y}_i^r(t)-\rho_i r,\tilde{Y}_i^r(t))\|\to 0, \tag{20}$$

and then

$$\sup_{t\in[0,Tr^{1-q}]} r^{-q}\|Z^r(t)-r\|\to 0. \tag{21}$$

In turn, to prove the latter statement it suffices to show that *there exists a construction of the underlying probability space, for which the statement holds.*

We will need some estimates, which can be obtained from a strong approximation of Poisson processes, available in, for example, [3, Chapters 1 and 2]:

PROPOSITION 19. *A unit rate Poisson process $\Pi(\cdot)$ and a standard Brownian motion $W(\cdot)$ can be constructed on a common probability space in such a way that the following holds. For some fixed positive constants C_1, C_2, C_3, such that $\forall T>1$ and $\forall u\ge 0$*

$$\mathbb{P}\left(\sup_{0\le t\le T}|\Pi(t)-t-W(t)|\ge C_1\log T+u\right)\le C_2e^{-C_3u}.$$

If in the above statement we replace T with rT, and u with $r^{1/4}$, we obtain

$$\mathbb{P}\left(\sup_{0\le t\le rT}|(\Pi(t)-t)-W(t)|<C_1\log(rT)+r^{1/4}\right)$$
$$>1-C_2e^{-C_3r^{1/4}}. \tag{22}$$

Note also that for a fixed $\delta\in(0,q-1/2)$ and all large r,

$$\mathbb{P}\left(\sup_{0\le t\le rT}|W(t)|\le r^{1/2+\delta}\right)\ge 1-e^{cr^{2\delta}} \tag{23}$$

for some constant $c>0$. If events in (22) and (23) hold for all large r, then

$$\sup_{0\le t\le rT} r^{-q}|\Pi(t)-t|\to 0. \tag{24}$$

To prove Theorem 18, consider the following construction of the probability space. (We want to strongly emphasize that this construction will be used only for the purpose of proving Theorem 18. For the proof of Theorem 7, we can and will use a different probability space construction.) For each r, we divide the time interval $[0,Tr^{1-q}]$ into r^{1-q} of T-long subintervals, namely $[(m-1)T,mT]$ with $m=1,2,\dots,r^{1-q}$. In each of the subintervals, and for each r, we consider independent unit rate Poisson processes $\Pi_i^{r,m}$, $\hat{\Pi}_i^{r,m}$, $\tilde{\Pi}_i^{r,m}$, driving type i exogenous arrivals, actual customer departures and token departures, respectively. More precisely, the number of type i exogenous arrivals, actual customer departures and token departures, by time t from the beginning of the m-th interval is given by

$$\Pi_i^{r,m}(\lambda_i rt),\quad \hat{\Pi}_i^{r,m}\left(\int_0^t\mu_i\hat{Y}_i^r(\xi)d\xi\right),\quad \tilde{\Pi}_i^{r,m}\left(\int_0^t\mu_0\tilde{Y}_i^r(\xi)d\xi\right),$$

respectively. Using (22)-(24) we obtain the following property for $\Pi_i^{r,m}$ (and analogous ones for $\hat{\Pi}_i^{r,m}$ and $\tilde{\Pi}_i^{r,m}$):

$$\max_{1\le m\le r^{1-q}}\max_{0\le t\le rT}|\Pi_i^{r,m}(t)-t|/r^q\to 0,\quad \text{as } r\to\infty,\quad \text{w.p.1.} \tag{25}$$

We denote

$$g^r(t)=(\hat{y}_i^r(t),\tilde{y}_i^r(t))=r^{-q}(\hat{Y}_i^r(t)-\rho_i r,\tilde{Y}_i^r(t)).$$

Then, we can prove the following.

LEMMA 20. *Consider fixed realizations (for each r) of driving processes, such that the properties (25) hold with q replaced by a smaller parameter $q'\in(1/2,q)$. Consider the corresponding sequence of realizations of $(g^r(t),\ t\ge 0)$, with bounded initial states $\|g^r(0)\|\le\epsilon$, $\epsilon>0$. Then, there exists a subsequence of r along which*

$$g^r(t)\to g(t),\quad u.o.c., \tag{26}$$

where $(g(t),\ t\ge 0)$ is Lipschitz continuous, with $\|g(0)\|\le\epsilon$, and it satisfies conditions

$$(d/dt)\hat{y}_i(t)=-\mu_i\hat{y}_i(t), \tag{27}$$

$$(d/dt)\tilde{y}_i(t)=\begin{cases}\mu_i\hat{y}_i(t)-\mu_0\tilde{y}_i(t), & \text{if } \tilde{y}_i(t)>0\\ \max\{0,\mu_i\hat{y}_i(t)-\mu_0\tilde{y}_i(t)\}, & \text{if } \tilde{y}_i(t)=0\end{cases} \tag{28}$$

at points $t\ge 0$, where the derivatives exist (which is almost everywhere w.r.t. the Lebesgue mesure). Moreover, the convergence

$$\|g(t)\|\to 0,\quad t\to\infty, \tag{29}$$

holds and is uniform w.r.t. initial states with $\|g(0)\|\le\epsilon$, and

$$\sup_{\|g(0)\|\le\epsilon}\max_{t\ge 0}\|g(t)\|\to 0,\quad \epsilon\to 0. \tag{30}$$

As a consequence of (30),

$$\|g(0)\|=0\ \ \textit{implies}\ \ \|g(t)\|=0,\ \forall t. \tag{31}$$

Lemma 20 is analogous to Lemma 14 in [14], except that the space scaling by r^{-q} is applied, as opposed to the fluid

scaling by r^{-1}, and the number of actual customers $\hat{Y}_i^r(t)$ is centered before scaling. The proof is somewhat more involved – the main issue is that (unlike for the fluid limit) the Lipschitz property of the limit is no longer automatic, because the rates of arrivals and departures in the system are $O(r)$, while the space is only scaled down by r^q. (That is why we need to use properties (25), as opposed to simply a strong law of large numbers.) However, this issue can be resolved as in, for example, the proof of Theorem 23 in [13]. We omit a detailed proof.

Proof of Theorem 18. By Corollary 17, we can choose a subsequence of r (increasing sufficiently fast) so that

$$\|g^r(0)\| \to 0, \quad \text{w.p.1.}$$

Then, we use the construction of the probability space specified above, which guarantees that w.p.1 the properties (25) hold with q replaced by a smaller parameter $q' \in (1/2, q)$ – let us consider any element of the probability space for which the properties (25) do hold. We claim that, for this element, (20) holds. Suppose not. Then, there exists $\epsilon > 0$ and a further subsequence of r, along which $\tau^r = \min\{t \mid \|g^r(t)\| > \epsilon\} \leq Tr^{1-q}$. By Lemma 20, we can and do choose time duration $T_1 > 0$ such that any limit trajectory $g(t)$ with $\|g(0)\| \leq \epsilon$ satisfies $\|g(T_1)\| \leq \epsilon/2$. For each r, consider the trajectory of g^r on the time interval $[\tau^r - T_1, \tau^r]$. (Suppose for now that $\tau^r \geq T_1$ for all sufficiently large r.) Then we can choose a further subsequence of r along which $g^r(\tau^r - T_1 + t) \to g(t)$ uniformly for $t \in [0, T_1]$, for a limit function $g(t)$ as in Lemma 20. But, this is impossible because then $\|g^r(\tau^r)\| \to \|g(T_1)\| \leq \epsilon/2$. The case when $\tau^r < T_1$ for infinitely many r is even simpler: we choose a further subsequence along which this is true, and consider the trajectories of g^r on the fixed time interval $[0, T_1]$. In this case any limit trajectory $g(t)$ described in Lemma 20 stays at 0 in the entire interval $[0, T_1]$, because $\|g(0)\| = \lim_r \|g^r(0)\| = 0$. This means that $\|g^r(\tau^r)\| \to 0$, again a contradiction. □

From this point on, we assume the following structure of the probability space. (It is different from the one used for the proof of Theorem 18, which, as we discussed, was for that proof only.) There are common (for all r) unit rate Poisson processes driving the system, defined as follows. For each $(\boldsymbol{k}, i) \in \mathcal{M}$ and $\hat{\boldsymbol{k}} \leq \boldsymbol{k}$, consider independent unit-rate Poisson process $\hat{\Pi}_{(\boldsymbol{k},\hat{\boldsymbol{k}}),i}(t),\ t \geq 0$, so that the number of actual type i customer departures from configuration $(\boldsymbol{k}, \hat{\boldsymbol{k}})$ in the interval $[0, t]$ is equal to $\hat{\Pi}_{(\boldsymbol{k},\hat{\boldsymbol{k}}),i}\left(\int_0^t \mu_i \hat{k}_i X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}(\xi) d\xi\right)$. Similarly, consider independent unit-rate Poisson process $\left\{\tilde{\Pi}_{(\boldsymbol{k},\hat{\boldsymbol{k}}),i}(t),\ t \geq 0\right\}$, so that the number of type i token departures from configuration $(\boldsymbol{k}, \hat{\boldsymbol{k}})$ due to their expiration, is equal to $\tilde{\Pi}_{(\boldsymbol{k},\hat{\boldsymbol{k}}),i}\left(\int_0^t \mu_0(k_i - \hat{k}_i) X^r_{(\boldsymbol{k},\hat{\boldsymbol{k}})}(\xi) d\xi\right)$. Finally, for each $i \in \mathcal{I}$, let $\{\Pi_i(t),\ t \geq 0\}$ be an independent unit-rate Poisson process, such that the number of exogenous type i arrivals in $[0, t]$ is equal to $\Pi_i(\lambda_i r t)$. For a fixed parameter $T > 0$, whose value will be chosen later, each of the above Poisson processes satisfies Lemma 12, in which we can and do replace T with $2T[(\bar{\mu} \vee \mu_0) + \sum_i \lambda_i]$. (We do this because we will "work" with system sample paths such that $\sum_i \hat{Y}_i = \sum_i (\hat{Y}_i^r + \tilde{Y}_i^r) < 2r$, and for these sample paths the total "instantaneous" rate of all transitions is upper bounded by $2r[(\bar{\mu} \vee \mu_0) + \sum_i \lambda_i]$.)

Denote by $\tilde{D}_i^r(t_1, t_2)$ the number of type-i token departures (due to their expirations), and by $\hat{A}_i^{**,r}(t_1, t_2)$ the total number of exogenous type-i arrivals (of actual customers) that do *not* replace type-i tokens, all in the interval $(t_1, t_2]$. Also, denote $Y_i^r(t_1, t_2) = Y_i^r(t_2) - Y_i^r(t_1)$.

THEOREM 21. *Consider the sequence (in r) of open systems in stationary regime. Let $T > 0$ be fixed. Then, any subsequence of r contains a further subsequence such that, w.p.1, the following holds:*

$$\tilde{D}_i^r(t_0, t_0 + r^{p-1})/[r^p r^{p-1}] \to 0, \tag{32}$$

$$\hat{A}_i^{**,r}(t_0, t_0 + r^{p-1})/[r^p r^{p-1}] \to 0, \tag{33}$$

uniformly on all intervals $[t_0, t_0 + r^{p-1}] \subset [0, Tr^{1-p}]$.

PROOF. Indeed, by Theorem 18, we can and do choose a subsequence of r along which (20)-(21) hold w.p.1. Then, (32) follows from (20), which states that the number of tokens $\tilde{Y}_i^r(t)$ is uniformly $o(r^p)$, and from the construction of the token departure processes, with the corresponding driving processes $\tilde{\Pi}_{(\boldsymbol{k},\hat{\boldsymbol{k}}),i}$ satisfying Lemma 12. From (20) we also have the uniform convergence

$$Y_i^r(t_0, t_0 + r^{p-1})/[r^p r^{p-1}] \to 0.$$

But, this along with (32) implies uniform convergence (33) as well, because we have the conservation law

$$Y_i^r(t_0, t_0 + r^{p-1}) = \hat{A}_i^{**,r}(t_0, t_0 + r^{p-1}) - \tilde{D}_i^r(t_0, t_0 + r^{p-1}).$$

The theorem is then proved. □

Proof of Theorem 7. Consider the sequence of the system processes in stationary regime. Consider a fixed $T > 0$, chosen to be sufficiently large, as in Proposition 15. Consider any subsequence of r. Then, we can and do choose a further subsequence of r along which, w.p.1, (20)-(21) hold with some $q \in (1/2, p)$ (by Theorem 18) , and the properties stated in Theorem 21 hold. As in the proof of Proposition 15, we will keep track of the evolution of the value of $F^r(\boldsymbol{X}^r(t))$. We emphasize that this is exactly the same function F^r as defined in Section 2.3 and used in the analysis of closed system, namely it has the fixed parameter r (in the system with index r), and *not* the random "parameter" Z^r. We claim that the following property holds.

Claim: *There exist positive constants $0 < C_1 < C_2$, $\delta > 0$, such that the following holds. For all sufficiently large r, uniformly on all intervals $[t_0, t_0 + r^{p-1}] \subset [0, Tr^{1-p}]$, we have (a) $F^r(\boldsymbol{X}^r(t_0)) - ru^* \geq C_1 r^p$ implies*

$$F^r(\boldsymbol{X}^r(t_0 + r^{p-1})) - F^r(\boldsymbol{X}^r(t_0)) \leq -\delta r^{2p-1},$$

and (b) $F^r(\boldsymbol{X}^r(t_0)) - ru^* \leq C_1 r^p$ implies

$$\sup_{\xi \in [0,1]} F^r(\boldsymbol{X}^r(t_0 + \xi r^{p-1})) - ru^* \leq C_2 r^p.$$

Clearly, (b) is analogous to Corollary 13 for the closed system and is proved exactly same way, with $\bar{\mu}$ in (14) replaced by $\bar{\mu} \vee \mu_0$. Statement (a) is analogous to Proposition 14 for the closed system, and we prove it below. It is also clear that the claim, along with (20)-(21), implies the theorem statement via the argument almost verbatim repeating that in the proof of Proposition 15.

It remains to prove (a). The proof is the same as that of Proposition 14, except that we have to make additional

estimates accounting for: (i) token departures due to their expiration and actual customer arrivals that do not find tokens; (ii) the fact that *GSS-M* uses weight function $\bar{w}^r = \bar{w}^r(X; Z^r)$, as opposed to function $w^r = w^r(X)$ (which has constant r as a parameter, instead of the random variable Z^r). This is because, *if we would have only transitions associated with actual customer departures and actual customer arrivals replacing tokens, and the assignment decisions would be based on weight w^r as opposed to $\bar{w}^r$*, then exactly the same drift estimates as those in the proof of Proposition 14 would apply. Note that in (i) we consider exactly those transitions for which we have properties (32)-(33). Therefore, in any interval $[t_0, t_0+r^{p-1}]$ the "worst case" possible increase in $F^r(\boldsymbol{X}^r)$ due to such transitions is $o(r^{2p-1})$. (We omit obvious epsilon/delta formalities.) Now consider (ii). Since we have the uniform bound $|Z^r(t) - r| \leq O(r^q)$, it is easy to check that $|\bar{w}^r(X) - w^r(X)| \leq O(r^{q-1})$ for any $X \geq 0$. This means that the error in the calculation of first-order contribution into the change of $F^r(\boldsymbol{X}^r)$ in any $[t_0, t_0 + r^{p-1}]$, introduced by *GSS-M* using weight $\bar{w}^r$ instead of w^r, is uniformly bounded by $O(rr^{p-1}r^{q-1}) = O(r^{p+q-1}) = o(r^{2p-1})$. (Again, we omit epsilon/delta formalities.) We see that the potential positive contribution of both (i) and (ii) into the change of objective function in any interval $[t_0, t_0 + r^{p-1}]$ is $o(r^{2p-1})$, uniformly on the choice of the interval. The estimate in (a) follows. Thus, the proof of the above claim, and of the theorem, follows. □

5. DISCUSSION

We presented the policy *Greedy with sublinear Safety Stocks (GSS)* along with a variant, which asymptotically minimize the steady-state total number of occupied servers at the fluid scale, as the input flow rates grow to infinity. A technical novelty of *GSS* is that it *automatically* creates non-zero safety stocks, *sublinear in the system "size"*, at server configurations which have zero stocks on the fluid scale. It is important to note that the algorithm does it without *a priori* knowledge of system parameters. To prove the fluid-scale optimality of *GSS*, we also need to consider a local fluid scaling, under which the sublinear safety stocks are "visible". This in turn allows us to obtain a tight asymptotic characterization of the algorithm deviation from exact optimal packing.

We can extend *GSS* to policies that asymptotically minimize the more general objective $\sum_{\boldsymbol{k}} c_{\boldsymbol{k}} X_{\boldsymbol{k}}$, where $c_{\boldsymbol{k}} > 0$ can be interpreted as the "cost" (for example, some estimated energy cost) of keeping a server in configuration $\boldsymbol{k}$, for each $\boldsymbol{k} \in \mathcal{K}$. Instead of the weight function $w^r(X^r_{\boldsymbol{k}})$ for each $\boldsymbol{k} \in \mathcal{K}$, consider the weight function $c_{\boldsymbol{k}} w^r(X^r_{\boldsymbol{k}})$, and define Δ^r as the difference between the new weight functions. We can then define *GSS* and *GSS-M* using the new Δ^r. They minimize the fluid scale quantity $\sum_{\boldsymbol{k}} c_{\boldsymbol{k}} x_{\boldsymbol{k}}$ asymptotically, and similar convergence rates can be obtained. If we assume that the cost $c_{\boldsymbol{k}}$ is monotonically non-decreasing in $\boldsymbol{k}$ (i.e., $c_{\boldsymbol{k}'} \leq c_{\boldsymbol{k}}$ if $\boldsymbol{k}' \leq \boldsymbol{k}$), then all our results and proofs still hold essentially verbatim. If costs $c_{\boldsymbol{k}}$ are not monotone in $\boldsymbol{k}$, most of the statements and proofs easily extend, except those of Lemmas 10 and 11, where some dual variables η_i may need to be negative. These η_i can be defined in a similar fashion as those in the proof of Lemma 6 in [14].

There are some possible directions for future research. For example, one may expect asymptotic optimality of "pure" *GSS* in an open system, which seems more difficult to establish. Proving or disproving its optimality may require better understanding of and some new insight into the system dynamics. Another direction can be the investigation of policies other (possibly simpler) than *GSS*. *GSS* is *asymptotically* optimal as the system scale increases. However, if the number $|\mathcal{K}|$ of feasible configurations is large, the system scale may need to be very large for the near optimal performance. It is then of interest to design policies (e.g., some form of best-fit) that have provably good performance properties at a wide range of system scales.

6. REFERENCES

[1] N. Bansal, A. Caprara, M. Sviridenko. A New Approximation Method for Set Covering Problems, with Applications to Multidimensional Bin Packing. *SIAM J. Comput.*, 2009, Vol.39, No.4, pp.1256-1278.

[2] J. Csirik, D. S. Johnson, C. Kenyon, J. B. Orlin, P. W. Shor, and R. R. Weber. On the Sum-of-Squares Algorithm for Bin Packing. *J.ACM*, 2006, Vol.53, pp.1-65.

[3] M. Csörgő and L. Horváth. *Weighted Approximations in Probability and Statistics*, Wiley, 1993.

[4] D. Gamarnik. Stochastic Bandwidth Packing Process: Stability Conditions via Lyapunov Function Technique. *Queueing Systems*, 2004, Vol.48, pp.339-363.

[5] D. Gamarnik and A. L. Stolyar. Multiclass Multiserver Queueing System in the Halfin–Whitt Heavy Traffic Regime: Asymptotics of the Stationary Distribution. *Queueing Systems*, 2012, Vol.71, pp.25-51.

[6] A. Gulati, A. Holler, M. Ji, G. Shanmuganathan, C. Waldspurger, and X. Zhu. VMware Distributed Resource Management: Design, Implementation and Lessons Learned. *VMware Technical Journal*, 2012, Vol.1, No.1, pp. 45-64. http://labs.vmware.com/publications/vmware-technical-journal

[7] V. Gupta and A. Radovanovic. Online Stochastic Bin Packing. Preprint, 2012.

[8] J. W. Jiang, T. Lan, S. Ha, M. Chen, and M. Chiang. Joint VM Placement and Routing for Data Center Traffic Engineering. *INFOCOM 2012*.

[9] S. T. Maguluri, R. Srikant, and L. Ying. Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters. *INFOCOM 2012*.

[10] S. Meyn. Dynamic Safety-Stocks for Asymptotic Optimality in Stochastic Networks. *Queueing Systems*, 2005, Vol. 50, pp.255-297.

[11] S. Shakkottai and A. L. Stolyar. Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule. *American Mathematical Society Translations*, 2002, Series 2, Vol. 207, pp. 185-202

[12] A. L. Stolyar and T. Tezcan. Shadow Routing Based Control of Flexible Multi-Server Pools in Overload. *Operations Research*, 2011, Vol.59, No.6, pp.1427-1444.

[13] A. L. Stolyar and E. Yudovina. Tightness of Invariant Distributions of a Large-Scale Flexible Service System under a Priority Discipline. Bell Labs Technical Memo, 2012. Submitted. http://arxiv.org/abs/1201.2978

[14] A. L. Stolyar. An Infinite Server System with General Packing Constraints. Bell Labs Technical Memo, 2012. Submitted. http://arxiv.org/abs/1205.4271

APPENDIX

A. PROOF OF LEMMA 3

Both $\mathcal{X}$ and $\mathcal{X}^*$ are convex and compact polytopes with a finite number of extreme points. Let $\mathcal{S}$ and $\mathcal{S}^*$ be the set of extreme points of $\mathcal{X}$ and of $\mathcal{X}^*$, respectively. Note that for all $\boldsymbol{x}^* \in \mathcal{S}^*$, $\sum_{\boldsymbol{k}} x^*_{\boldsymbol{k}} = u^*$, and for all $\boldsymbol{x}' \in \mathcal{S}\backslash\mathcal{S}^*$, $\sum_{\boldsymbol{k}} x'_{\boldsymbol{k}} > u^* + \delta$. for some $\delta > 0$.

Let $\langle \mathcal{S}\backslash\mathcal{S}^* \rangle$ be the convex hull of the set $\mathcal{S}\backslash\mathcal{S}^*$. Then for all $\boldsymbol{x}' \in \langle \mathcal{S}\backslash\mathcal{S}^* \rangle$, $\sum_{\boldsymbol{k}} x'_{\boldsymbol{k}} \geq u^* + \delta$. Consider the function $g : \langle \mathcal{S}\backslash\mathcal{S}^* \rangle \times \mathcal{X}^* \to \mathbb{R}$ defined by $g(\boldsymbol{x}^*, \boldsymbol{x}') = \|\boldsymbol{x}^* - \boldsymbol{x}'\| / \left(\sum_{\boldsymbol{k}\in\mathcal{K}} x'_{\boldsymbol{k}} - u^*\right)$. Function g is well-defined, always positive and clearly continuous. Since both $\langle \mathcal{S}\backslash\mathcal{S}^* \rangle$ and $\mathcal{X}^*$ are compact, so is their product space. Thus there exists $D > 0$ such that g is upper bounded by D.

For every $\boldsymbol{x} \in \mathcal{X}$, there exists $\lambda \in [0, 1]$ such that $\boldsymbol{x} = \lambda \boldsymbol{x}' + (1-\lambda)\boldsymbol{x}^*$, with $\boldsymbol{x}' \in \langle \mathcal{S}\backslash\mathcal{S}^* \rangle$ and $\boldsymbol{x}^* \in \mathcal{X}^*$. Then

$$\begin{aligned} d(\boldsymbol{x}, \mathcal{X}^*) &\leq \|\boldsymbol{x} - \boldsymbol{x}^*\| = \lambda \|\boldsymbol{x}' - \boldsymbol{x}^*\| \\ &= \lambda g(\boldsymbol{x}^*, \boldsymbol{x}') \left(\sum_{\boldsymbol{k}\in\mathcal{K}} x'_{\boldsymbol{k}} - u^*\right) \\ &\leq \lambda D \left(\sum_{\boldsymbol{k}\in\mathcal{K}} x'_{\boldsymbol{k}} - u^*\right) = D \left(\sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} - u^*\right). \end{aligned}$$

B. PROOF OF PROPOSITION 15

Let $\delta > 0$ be the same as in Proposition 14, and define $T = 3/\delta$. $C > 0$ will be chosen to be sufficiently large, whose value will be determined later in the proof. Clearly, to prove the proposition, it suffices to prove a stronger property

$$\mathbb{P}\left(d\left(\boldsymbol{x}^r(Tr^{1-p}), \mathcal{X}^*\right) \leq Cr^{p-1} \text{ for all large } r\right) = 1.$$

By Proposition 14, there exists $C_1 > 0$ such that w.p.1, for sufficiently large r, and for any interval $[t_0, t_0 + r^{p-1}] \subset [0, Tr^{1-p}]$, if $d\left(\boldsymbol{x}^r(t_0), \mathcal{X}^*\right) \geq C_1 r^{p-1}$, then

$$F^r\left(\boldsymbol{X}^r(t_0^r + r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(t_0^r)\right) \leq -\delta r^{2p-1}. \tag{34}$$

We pick some r such that the above statement holds, and that furthermore, for every $t_0 \in [0, Tr^{1-p}]$ and $\xi \in [0, 1]$,

$$d\left(\boldsymbol{X}^r(t_0^r + \xi r^{p-1}), \boldsymbol{X}^r(t_0^r)\right) \leq O(r^p). \tag{35}$$

This can be done by Corollary 13.

Now claim that $d(\boldsymbol{x}^r(Tr^{1-p}), \mathcal{X}^*) \leq Cr^{p-1}$. To establish the claim, we consider the set $\mathcal{L} = \{\ell \in \mathbb{Z}_+ : \ell r^{p-1} \in [0, Tr^{1-p}]\}$, and prove that

(a) there exists $\ell_0 \in \mathcal{L}$ such that $d(\boldsymbol{x}^r(\ell_0 r^{p-1}), \mathcal{X}^*) \leq C_1 r^{p-1}$, and

(b) there exists $C_2 > 0$ such that for all $\ell \in \mathcal{L}$ with $\ell \geq \ell_0$, $F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq ru^* + C_2 r^p$.

First suppose that (a) does not hold. Then for every $\ell \in \mathcal{L}$, $d(\boldsymbol{x}^r(\ell r^{p-1}), \mathcal{X}^*) \geq C_1 r^{p-1}$, so

$$F^r\left(\boldsymbol{X}^r((\ell+1)r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq -\delta r^{2p-1}.$$

Let $\bar{\ell} = \lceil Tr^{2(1-p)} \rceil$. Summing these inequalities over ℓ, we obtain

$$\begin{aligned} &F^r\left(\boldsymbol{X}^r(\bar{\ell} r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(0)\right) \leq -\bar{\ell}\delta r^{2p-1} \\ &\leq -(Tr^{2(1-p)} - 1)\delta r^{2p-1} = -T\delta r + \delta r^{2p-1}. \end{aligned}$$

Thus,

$$\begin{aligned} &F^r\left(\boldsymbol{X}^r(\bar{\ell} r^{p-1})\right) \leq F^r\left(\boldsymbol{X}^r(0)\right) - T\delta r + \delta r^{2p-1} \\ &\leq r - \frac{3}{\delta}\delta r + \delta r^{2p-1} < 0. \end{aligned}$$

This contradicts the nonnegativity of F^r, so statement (a) is established.

To establish statement (b), we use the following simple lemma, whose proof is omitted.

LEMMA 22. *Let K, α and β be given positive constants. Consider a sequence of real numbers $\{a_n\}$ that satisfies: (i) $a_0 \leq K$, (ii) $a_{n+1} - a_n \leq \alpha$, and (iii) if $a_n \geq K$, then $a_{n+1} - a_n \leq -\beta$. Then $\max_n a_n \leq K + \alpha$.*

We will establish the following corresponding statements:

(i) $F^r\left(\boldsymbol{X}^r(\ell_0 r^{p-1})\right) \leq ru^* + C_1 r^p$. Recall that we have $d\left(\boldsymbol{x}^r(\ell_0 r^{p-1}), \mathcal{X}^*\right) \leq C_1 r^{p-1}$, so by Lemma 4,

$$\begin{aligned} &F^r\left(\boldsymbol{X}^r(\ell_0 r^{p-1})\right) - ru^* \leq \sum_{\boldsymbol{k}\in\mathcal{K}} X^r_{\boldsymbol{k}}(\ell_0 r^{p-1}) - ru^* \\ &\leq rd\left(\boldsymbol{x}^r(\ell_0 r^{p-1}), \mathcal{X}^*\right) \leq C_1 r^p. \end{aligned}$$

(ii) There exists $C_3 > 0$ such that $F^r\left(\boldsymbol{X}^r((\ell+1)r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq C_3 r^p$. This is clear, since by Lemma 4, $F^r\left(\boldsymbol{X}^r\right)$ differs from $\sum_{\boldsymbol{k}} X^r_{\boldsymbol{k}}$ by $O(r^p)$, and the change in $\boldsymbol{X}^r$ is at most $O(r^p)$ over an interval of length r^{1-p}.

(iii) If $F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \geq ru^* + C_1 r^p$, then

$$F^r\left(\boldsymbol{X}^r((\ell+1)r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq -\delta r^{2p-1}.$$

To see this, suppose that $F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \geq ru^* + C_1 r^p$. Then $d\left(\boldsymbol{x}^r(\ell r^{p-1}), \mathcal{X}^*\right) \geq \sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}} - u^* \geq \frac{1}{r} F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) - u^* \geq C_1 r^{p-1}$, and we must have

$$F^r\left(\boldsymbol{X}^r((\ell+1)r^{p-1})\right) - F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq -\delta r^{2p-1}.$$

By Lemma 22, for all $\ell \in \mathcal{L}$ with $\ell \geq \ell_0$, we have

$$F^r\left(\boldsymbol{X}^r(\ell r^{p-1})\right) \leq ru^* + (C_1 + C_3)\, r^p = ru^* + C_2 r^p,$$

by letting $C_2 = C_1 + C_3$. This establishes statement (b). In particular, for $\bar{\ell} = \lceil Tr^{2(1-p)} \rceil$,

$$F^r\left(\boldsymbol{X}^r(\bar{\ell} r^{p-1})\right) \leq ru^* + C_2 r^p.$$

Now by (35), the difference between $\boldsymbol{X}^r(Tr^{1-p})$ and $\boldsymbol{X}^r(\bar{\ell} r^{p-1})$ is $O(r^p)$. Furthermore, the difference between $F^r\left(\boldsymbol{X}^r(\bar{\ell} r^{p-1})\right)$ and $\boldsymbol{X}^r(\bar{\ell} r^{p-1})$ also $O(r^p)$. This implies that

$$\sum_{\boldsymbol{k}\in\mathcal{K}} X^r_{\boldsymbol{k}}(Tr^{1-p}) - ru^* \leq C_2 r^p + O(r^p).$$

Thus, there exists $C > 0$ such that

$$\sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}}(Tr^{1-p}) - u^* \leq \frac{C}{D} r^{p-1}.$$

By Lemma 3,

$$\begin{aligned} &d(\boldsymbol{x}^r(Tr^{1-p}), \mathcal{X}^*) \leq D \left(\sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}}(Tr^{1-p}) - u^*\right) \\ &\leq D \cdot \frac{C}{D} r^{1-p} = Cr^{1-p}, \end{aligned}$$

and we have established the claim. Therefore, w.p.1,

$$d(\boldsymbol{x}^r(Tr^{1-p}), \mathcal{X}^*) \leq Cr^{1-p},$$

for all sufficiently large r. This establishes the proposition.

C. PROOF OF THEOREM 16

The general approach of the proof is similar to that of Theorem 2 (ii) in [5], in that it is based on the process generator estimates for the exponent e^{Φ}, where Φ is a function on the state space. However, the function Φ in our case is much different, and so are the specifics of the estimates. Consider fixed $i \in \mathcal{I}$ and r. For notational convenience, we drop the subscript i and superscript r from all quantities considered in this proof. The Markov chain $\boldsymbol{U}(\cdot) = (\hat{Y}(\cdot), \tilde{Y}(\cdot))$ has infinitesimal transition rate matrix $\boldsymbol{\xi}$ given by

$$\xi(\boldsymbol{u}, \boldsymbol{u}+\boldsymbol{v}) \to \begin{cases} \lambda r, & \text{if } \boldsymbol{v} = (1, -1 \cdot \mathbf{1}_{\{\tilde{y}>0\}}), \\ \mu\hat{y}, & \text{if } \boldsymbol{v} = (-1, 1), \\ \mu_0\tilde{y}, & \text{if } \boldsymbol{v} = (0, -1), \\ 0, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{u} = (\hat{y}, \tilde{y})$. We consider A the infinitesimal generator of the Markov chain $\boldsymbol{U}(\cdot)$, defined by

$$AG(\boldsymbol{u}) = \sum_{\boldsymbol{u}'} \xi(\boldsymbol{u}, \boldsymbol{u}') \left(G(\boldsymbol{u}') - G(\boldsymbol{u})\right), \tag{36}$$

for all functions $G : \mathbb{Z}_+^2 \to \mathbb{R}$ in the domain of A. We also consider the formal operator $\bar{A}$, defined (similar to Eq. (36)) by

$$\bar{A}G(\boldsymbol{u}) = \sum_{\boldsymbol{u}'} \xi(\boldsymbol{u}, \boldsymbol{u}') \left(G(\boldsymbol{u}') - G(\boldsymbol{u})\right), \tag{37}$$

for all functions $G : \mathbb{Z}_+^2 \to \mathbb{R}$. Similarly to [5], it is easy to observe that the following property holds: if a function G takes a fixed constant value on the entire state space, except maybe a finite subset, then G is within the domain of A, $AG = \bar{A}G$, and moreover

$$\mathbb{E}[AG(\boldsymbol{U})] = \mathbb{E}[\bar{A}G(\boldsymbol{U})] = 0, \tag{38}$$

where the expectation is taken w.r.t the stationary distribution of the Markov chain $\boldsymbol{U}(\cdot)$.

First, define the (candidate) Lyapunov function $G : \mathbb{Z}_+^2 \to \mathbb{R}$ by

$$G(\boldsymbol{u}) = \exp\left(\frac{1}{\sqrt{r}} h(\boldsymbol{u})\right),$$

where $h(\boldsymbol{u}) = \sqrt{(\hat{y} - \rho r)^2 + \frac{\mu_0}{\mu}\tilde{y}^2}$. Note that, for an arbitrary $b \geq 0$, the truncated function

$$\bar{G}^{(b)}(\boldsymbol{u}) = \exp\left(\frac{h(\boldsymbol{u})}{\sqrt{r}} \wedge b\right)$$

is constant outside a finite subset and therefore, by (38),

$$\mathbb{E}[\bar{A}G^{(b)}(\boldsymbol{U})] = 0. \tag{39}$$

Also note that,

$$\bar{A}G^{(b)}(\boldsymbol{u}) \leq \bar{A}G(\boldsymbol{u}), \quad \text{if } h(\boldsymbol{u})/\sqrt{r} \leq b,$$

$$\bar{A}G^{(b)}(\boldsymbol{u}) \leq 0, \quad \text{if } h(\boldsymbol{u})/\sqrt{r} \geq b.$$

Similar to [5], the following inequality can be derived, using Taylor expansion. There exists some constant $c_2 > 0$ such that for sufficiently large r,

$$\bar{A}G(\boldsymbol{u}) \leq G(\boldsymbol{u}) \left(\frac{1}{\sqrt{r}} \bar{A}h(\boldsymbol{u}) + \frac{c_2}{r}(\lambda r + \mu\hat{y} + \mu_0\tilde{y})\right). \tag{40}$$

The term $\frac{G(\boldsymbol{u})}{\sqrt{r}} \bar{A}h(\boldsymbol{u})$ captures the first-order change in $G(\boldsymbol{u})$, and $\frac{c_2 G(\boldsymbol{u})}{r}(\lambda r + \mu\hat{y} + \mu_0\tilde{y})$ bounds the second-order change. Here we used the fact that h is Lipschitz continuous and $\|\boldsymbol{u}\|$ is changed by at most 1 by any single transition. Now consider the term $\bar{A}h(\boldsymbol{u})$. We use the following inequality to bound $\bar{A}h(\boldsymbol{u})$:

$$\sqrt{(x+a)^2 + (y+b)^2} - \sqrt{x^2 + y^2} \leq \frac{ax + by + a^2 + b^2}{\sqrt{x^2 + y^2}}.$$

To verify this inequality, note that first,

$$\left(\sqrt{(x+a)^2 + (y+b)^2}\right)^2 \leq \left(\sqrt{x^2 + y^2} + \frac{ax + by + a^2 + b^2}{\sqrt{x^2 + y^2}}\right)^2,$$

and second,

$$\sqrt{x^2 + y^2} + \frac{ax + by + a^2 + b^2}{\sqrt{x^2 + y^2}} \geq 0.$$

Thus,

$$\begin{aligned} \bar{A}h(\boldsymbol{u}) &\leq \frac{(\lambda r - \mu\hat{y})(\hat{y} - \rho r) - (\lambda r - \mu\hat{y} + \mu_0\tilde{y})(\mu_0\tilde{y}/\mu)}{\sqrt{(\hat{y} - \rho r)^2 + \mu_0\tilde{y}^2/\mu}} \\ &\quad + \frac{c_3(\lambda r + \mu\hat{y} + \mu_0\tilde{y})}{\sqrt{(\hat{y} - \rho r)^2 + \mu_0\tilde{y}^2/\mu}} \\ &= \frac{-\frac{\mu}{2}(\hat{y} - \rho r)^2 - \frac{\mu_0^2}{2\mu}\tilde{y}^2 - \frac{\mu}{2}(\hat{y} - \rho r + \tilde{y})^2}{\sqrt{(\hat{y} - \rho r)^2 + \mu_0\tilde{y}^2/\mu}} \\ &\quad + \frac{c_3(\lambda r + \mu\hat{y} + \mu_0\tilde{y})}{\sqrt{(\hat{y} - \rho r)^2 + \mu_0\tilde{y}^2/\mu}} \\ &\leq \frac{-\frac{\mu}{2}(\hat{y} - \rho r)^2 - \frac{\mu_0^2}{2\mu}\tilde{y}^2}{h(\boldsymbol{u})} + \frac{c_3(\lambda r + \mu\hat{y} + \mu_0\tilde{y})}{h(\boldsymbol{u})} \\ &\leq -c_4 h(\boldsymbol{u}) + \frac{c_3}{\sqrt{r}}(\lambda r + \mu\hat{y} + \mu_0\tilde{y}), \end{aligned} \tag{41}$$

for some positive constants c_3 and c_4, and when $h(\boldsymbol{u}) \geq \sqrt{r}$. Combining Inequalities (40) and (41), we have

$$\bar{A}G(\boldsymbol{u}) \leq G(\boldsymbol{u}) \left(-\frac{c_4}{\sqrt{r}} h(\boldsymbol{u}) + \frac{c_2 + c_3}{r}(\lambda r + \mu\hat{y} + \mu_0\tilde{y})\right).$$

Consider the term in the bracket on the RHS. It is now an elementary calculation to see that there exists some positive constant c_5, such that whenever $h(\boldsymbol{u}) \geq c_5\sqrt{r}$,

$$-\frac{c_4}{\sqrt{r}} h(\boldsymbol{u}) + \frac{c_2 + c_3}{r}(\lambda r + \mu\hat{y} + \mu_0\tilde{y}) \leq -1.$$

Also note that when $h(\boldsymbol{u}) < c_5\sqrt{r}$, the maximum values of

$$G(\boldsymbol{u}) \quad \text{and} \quad G(\boldsymbol{u}) \left(-\frac{c_4}{\sqrt{r}} h(\boldsymbol{u}) + \frac{c_2 + c_3}{r}(\lambda r + \mu\hat{y} + \mu_0\tilde{y})\right)$$

are both bounded above by an absolute constant, say c_6, which does not depend on r. In summary,

$$\bar{A}G(\boldsymbol{u}) \leq -G(\boldsymbol{u}) \quad \text{whenever} \quad h(\boldsymbol{u}) \geq c_5\sqrt{r},$$
$$\text{and} \quad \bar{A}G(\boldsymbol{u}) \leq c_6 \quad \text{whenever} \quad h(\boldsymbol{u}) < c_5\sqrt{r}.$$

Thus, for any $b > c_5$,

$$\begin{aligned} 0 = \mathbb{E}[AG^{(b)}(\boldsymbol{U})] &\leq \mathbb{E}[\bar{A}G(\boldsymbol{U})\mathbf{1}_{\{c_5\sqrt{r} \leq h(\boldsymbol{U}) \leq b\sqrt{r}\}}] \\ &\quad + \mathbb{E}[\bar{A}G(\boldsymbol{U})\mathbf{1}_{\{h(\boldsymbol{U}) < c_5\sqrt{r}\}}] \\ &\leq -\mathbb{E}[G(\boldsymbol{U})\mathbf{1}_{\{c_5\sqrt{r} \leq h(\boldsymbol{U}) \leq b\sqrt{r}\}}] + c_6. \end{aligned}$$

This implies that $\mathbb{E}[G(\boldsymbol{U})\mathbf{1}_{\{c_5\sqrt{r} \leq h(\boldsymbol{U}) \leq b\sqrt{r}\}}] \leq c_6$, and then $\mathbb{E}[G(\boldsymbol{U})\mathbf{1}_{\{h(\boldsymbol{U}) \leq b\sqrt{r}\}}] \leq 2c_6$. Finally, by Monotone Convergence, $\mathbb{E}[G(\boldsymbol{U})] \leq 2c_6$. This completes the proof.